

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2021

Limits of Bayesianism

Zhang, Xueyin (Snow)

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp016969z401v>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

LIMITS OF BAYESIANISM

Xueyin Zhang

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
PHILOSOPHY

Advisers: Adam Elga and Harvey Lederman

September 2021

© Copyright by Xueyin Zhang, 2021.

All rights reserved.

Abstract

Two central questions in epistemology are: What should rational agents believe? How should they revise their beliefs in light of new information? Bayesianism offers a unified answer to both. Roughly, the view consists of two claims:

Probabilistic Coherence. Rational agents have degrees of belief (credences) that can be represented by a real-valued probability function over an algebra of propositions.

Conditionalization. Rational agents update their credences by conditionalizing on their total evidence.

The view appears simple and powerful. It explains, for instance, why it seems irrational to be more confident that there are aliens on Mars than that there are aliens, or to change one's mind about the existence of Martians after learning that snow is white, if one judges the two to be unconnected *ex ante*. The former violates the axioms of probability, while the latter violates the rule of conditionalization.

This dissertation examines the limits of Bayesianism from within its own theoretical confines. I look at three independent but related questions:

1. Should rational agents have degrees of belief that are countably additive?
2. Should rational agents refrain from having definite degrees of belief given certain propositions?
3. How should rational agents plan to update in response to new information in general?

I show that, in each case, the naive Bayesian answer conflicts with some intuitively plausible principles that many Bayesians endorse on independent grounds. The upshot is not purely negative. I argue that these conflicts shed interesting light on general epistemological questions, including the role of infinitary idealization, the nature of evidence and the value of rationality.

Acknowledgments

In writing this dissertation, I benefited tremendously from stimulating conversations and exchanges with many individuals, including David Builes, Eliya Cohen, Kenny Easwaran, Dmitri Gallow, Hans Halvorson, Grace Helton, Daniel Hoek, Kyle Landrum, Stephen Mackereth, Sarah McGrath, Calum McNamara, Chris Register, Alejandro Naranjo Sandoval, Aarthy Vaidyanathan and Erik Zhang. (Sincere apologies to anyone whom I have inadvertently missed.) Special thanks are due to my readers, Mark Johnston and Boris Kment, for reading the dissertation on short notice and for providing very detailed and constructive feedback.

I would like to thank the following individuals in particular:

To my advisors, Adam Elga and Harvey Lederman, for their unwavering support and invaluable guidance. It has been truly a privilege to work with both of them in the past few years. They somehow managed to find faith in me when I struggled to find it in myself. Without them, this dissertation would not have existed, and I probably would not have been a philosopher.

More specifically: Thanks to Adam for showing me how to do philosophy in a way that is formally rigorous but also fun and exciting. It is a skill that I have yet to fully master. Much of the dissertation grew out of my own struggle with either Adam's work or some penetrating (often monosyllabic) remarks he made in our conversations. I expect that struggle to continue, and I can't wait to see where it will lead me next.

Thanks to Harvey for being always so encouraging, compassionate and extremely generous with his time. The ideas in this dissertation have become so much better, thanks to his extensive, constructive, and incredibly prompt feedback. There is a Chinese proverb that says "A single conversation across the table with a wise person is better than a ten-year study of books." It captures my feeling every time I've had a conversation with Harvey.

To my unofficial advisor, Michael Rescorla, for his invaluable mentorship and extremely helpful feedback on early drafts of chapters of this dissertation as well as my joint works with Alexander Meehan. Much of the formal framework I've developed in this dissertation is indebted to Michael's pioneering works on Dutch books and regular conditional distributions, which continue to be a constant source of inspiration.

To Alexander Meehan, for being the most wonderful collaborator and intellectual interlocutor. Both

Chapter 3 and Chapter 5 were inspired by our joint works on related topics. Alex has taught me by example how to approach philosophy with care, rigor and open-mindedness. Working with Alex has been one of the best highlights of my graduate school experience, and hopefully our collaboration continues on for the many years to come.

To Claudia Yau, my philosophical twin, for her unreserved support and care. Talking with her, about philosophy or life in general, has always been an absolute delight. It has been such a great privilege to complete the last ten years of my philosophical education in the company of such a great mind and gentle soul, and I am confident that our paths will cross once again in the near future.

To my Wellesley professors, in particular Catherine Wearing, Corinne Gartner, Julie Walsh and Stanley Chang, for their invaluable mentorship and sage advice. They went above and beyond to help and support me even after I graduated, and I am deeply, deeply grateful.

Lastly, to my parents, for their unconditional love and support. Their virtual company was what got me through the most difficult periods of graduate school. I wish I had spent more time with them in the past few years. I love them 3000.

This dissertation is dedicated to the memories of my grandfather. I know he could not have cared less about whether I got a PhD. All he had ever cared were my health and happiness. I miss him very, very much.

Contents

1	Conglomerability and Countable Additivity	10
1	Introduction	12
2	Preliminaries	12
3	The conglomerability argument	16
4	Begging the question	17
5	Question-sensitivity	19
6	Credal states	23
7	Coda	33
2	Closure and Countable Additivity	36
1	Introduction	36
2	Closure	37
3	First argument	41
4	Second argument	46
5	Replies	49
6	A way out	53
3	Borel and Bertrand	55
1	The paradox	56
2	Relativism	59
3	Critique of Relativism	60
4	Borel and Bertrand	73
5	Conclusion	77
4	Conditioning and Crossing the Bridge	79
1	Introduction	79

2	Self-Stability	81
3	Luminosity	88
4	Immodesty	92
5	Reflection	95
6	Crossing the bridge	98
5	Deference and Decisions	102
1	Introduction	102
2	Setup	103
3	First Result	106
4	Second Result	116
5	Sensitivity and rationality	117
	Appendices	129
A	Chapter 1	130
1	Global Coherence	130
2	Freiling and Lévy	131
3	Builes' paradox	133
B	Chapter 2	137
C	Chapter 4	142
D	Chapter 5	145

Introduction

There are two central questions in epistemology: What should rational agents believe? How should they revise their beliefs in light of new information? Bayesianism offers a unified answer to both. Roughly, the view consists of two claims:

Probabilistic Coherence. Rational agents should have degrees of belief (i.e. credences) that can be represented by a real-valued probability function over an algebra of propositions.¹

Conditionalization. Rational agents should update their credences by conditionalizing on their total evidence.

The view appears simple and powerful. It explains, for instance, why it would be irrational to be more confident that there are aliens on Mars than that there are aliens, or to change one's mind about whether there are aliens on Mars after learning that snow is white (provided that one takes the two to be irrelevant). The former violates the axioms of probability, while the latter violates the rule of conditionalization.

Despite its success, Bayesianism has its problems. On the one hand, the view seems too permissive. An agent can satisfy the two Bayesian requirements and yet have credences that seem intuitively irrational (e.g. they can be maximally confident that the Moon is made of blue cheese). On the other hand, the view also seems too stringent. It follows from Bayesianism that rational agents should be logically omniscient. Yet intuitively, someone who knows basic arithmetic and calendar facts can be rationally uncertain about what date is 1234567 days after April 6, 2021,² even though the answer is entailed by what they know. For these reasons, theorists have either rejected the framework (Weisberg 2015b), or tried to enrich it with additional principles (e.g. the Principal Principle, Jeffrey conditionalization) or representational tools (e.g. imprecise probabilities, infinitesimal probabilities, inquisitive probabilities).

¹Two qualifications. First, about objects of credences: Throughout this dissertation, I shall take objects of credences to be propositions while staying neutral on what they are (sets of possible worlds, structured entities, sets of impossible worlds, etc.). Second, Probabilistic Coherence as stated is incompatible with the view that rational agents have degrees of beliefs that are only representable by sets of probability functions, or probability functions whose ranges have more complicated mathematical structures than the reals (e.g. hyperreals). I shall be focusing on this stronger version of Probabilistic Coherence mainly for the sake of simplicity, though many of my arguments in Chapter 1-3 extend to the weaker version(s) of Probabilistic Coherence as well.

²Answer: June 21, 340034.

This dissertation examines Bayesianism from a different perspective.³ Instead of looking at the framework “from the outside,” I raise problems for the view from within its own theoretical confines. More specifically, I show that certain principles/perspectives that seem implicit in the Bayesian framework turn out to be jointly incompatible with each other. The upshot is not purely negative. I argue that these conflicts also shed interesting light on general epistemological questions, such as the role of infinitary idealization, the nature of evidence and confirmation, and the value of rationality.

Plan. This dissertation divides into two parts. The first part (Chapter 1-3) discusses two related problems: (i) the normative status of countable additivity and (ii) the Borel-Kolmogorov paradox. More specifically: Chapter 1 argues that one of the strongest arguments in favor of the axiom of countable additivity, the conglomerability argument, does not succeed under scrutiny. Chapter 2 shows that the axiom is in conflict with some fairly weak epistemological principles, many of which the Bayesian accepts independently. Taken together, these two chapters cast doubt on the idea that countable additivity is a universal rational norm.

Some of my arguments in the first two chapters rely on the theory of coherent conditional probability. Some Bayesians reject that theory on the basis of the Borel-Kolmogorov paradox. In Chapter 3, I argue that this is a mistake; instead, the Borel-Kolmogorov paradox should be understood as an instance of the problem of priors. As a result, there is indeed a tension between holding countable additivity as a norm of coherence on the one hand, and rejecting the coherent theory of conditional probability on the basis of Borel-Kolmogorov paradox on the other.

The second part (Chapter 4 and 5) changes gears and starts out with a putative challenge to Bayesianism: In some cases where the agent lacks introspective access to her evidence and/or her own mental state, Conditionalization recommends belief changes that seem rationally defective. In Chapter 4, I offer a diagnosis of this rational defect: Conditionalization is irrational in such situations as an *updating plan*. I propose a new norm for rational plans (including updating plans), *Self-Stability*, and discuss how it relates to other existing epistemological concepts such as immodesty, luminosity and deference-worthiness. Lastly, in Chapter 5, I show that what appeared as a conflict between Bayesianism and Self-Stability turns out to be a deeper conflict between Bayesianism and cognitive decision theory – the view that updating plans should be evaluated as options in a decision problem. How we should adjudicate this conflict is a question that I shall leave for future investigation.

Before proceeding, let me make one preliminary remark about Conditionalization. The reader can skip

³Why beat up a horse that is already dead? The short answer is that Bayesianism is not dead. Indeed, quite the contrary, the view seems very much to be the received orthodoxy in contemporary epistemology. I think this has partly to do with the logical structure of the existing objections to Bayesians. Most of these objections are arguments by counter-examples. As a result, it is always open to the Bayesians to respond by restricting the scope of their theory. For instance, one popular reply to the problem of logical omniscience is that Bayesianism is a theory of ideal rationality; it only applies to agents with unlimited cognitive resources and computational power (cf. Stalnaker 1991; Christensen 1991; a similar move is made by Lewis in response to Jeffrey’s objection that Bayesianism fails to accommodate learning with uncertainties).

this section and refer back to it when necessary.

Conditionalization

While Probabilistic Coherence seems relatively straightforward (bracketing the complications related to logical omniscience), the content of Conditionalization is a lot less clear. Roughly, it says that rational agents should update by conditionalizing on their total evidence. But what *is* conditionalization? More specifically, given one's prior probability function P and total evidence E , what *is* the conditional probability function $P(.|E)$?

In one sense, the answer is simple: When the prior probability of E is positive, everyone agrees that the conditional probability function $P(.|E)$ can be calculated by the ratio formula:

$$P(.|E) = \frac{P(. \& E)}{P(E)}.$$

However, we can't use this formula when $P(E) = 0$, for in this case the right-hand side is undefined. In particular, if $P(E) = 0$, then by finite additivity, $P(A \& E) = 0$ for any A in the domain of P . It follows that, when $P(E) = 0$, the value of $P(A \& E)/P(E)$ can't be a function of $P(A \& E)$ and $P(E)$. For if there were such a function (call it " f "), then $P(A|E)$ would equal $f(0, 0)$ for all A . But this means that $P(.|E)$ would have to be a constant function and therefore probabilistically incoherent.

The above argument shows that $P(.|E)$ cannot be fully determined by the unconditional probability function P when $P(E) = 0$; additional information is required. Almost everyone agrees with this point. However, the agreement stops pretty much there. In particular, theorists diverge on the following two questions:

1. Is $P(.|E)$ always well-defineable when $P(E) = 0$?
2. Suppose $P(.|E)$ is well-defineable. What, if any, information is required in order to specify $P(.|E)$?
(Or, in the normative context, what rational constraints are there on $P(.|E)$?)

Both questions will come up again in the following chapters. For present purposes, it suffices to note that the answer one gives to either question may affect how one precisifies the content of Conditionalization. In particular, there are at least three possible options:

1. Conditionalization is silent when the agent assigns zero prior credence to what she learns.
2. Conditionalization does not distinguish between null updating and normal updating. The agent is assumed to have well-defined conditional credences in A given B , for any proposition A and non-

contradictory proposition B . To conditionalize on B is to adopt the credence function $P_{old}(.|B)$, where $P_{old}(.|.)$ is the agent's prior (primitive) conditional credence function.

3. Conditionalization is not silent when the agent learns zero-probability evidence, but does treat those cases differently than normal updating.

To avoid terminological confusion, I shall adopt the second interpretation of Conditionalization for the rest of this dissertation. That is, I take Conditionalization to say that: if an agent's prior conditional credence in A given B is x , then the agent's posterior conditional credence in A after learning B and nothing else should also be x .

Having settled the question of conditional probability, let me turn now to a different question – or rather, five of them:

1. *Diachronic vs. Plan.* Is Conditionalization a diachronic constraint on rational posteriors, or a synchronic constraint on rational updating plans?

I shall say a lot more about updating plans in the following chapters. So let me be brief here. On a very intuitive level, an updating plan specifies what credence function the agent will adopt given some/all possible evidence that she might learn at a future time.⁴ Clearly, there is an intimate relationship between norms for rational posteriors and norms for rational updating plans. In particular, the latter entails the former if we assume *Diachronic Continence* (Paul 2014), the principle that says, everything equal, one should update in accord with one's plan. On the other hand, despite its intuitive plausibility, Diachronic Continence has been disputed in the literature (cf. Pettigrew (2016a)), and to this extent it is worth keeping the distinction in place.

2. *Wide vs. Narrow Scope.* Is Conditionalization a narrow-scope norm on one's posterior/updating plan (given one's prior and learning experience) or a wide-scope norm on the combination of one's prior and posterior/updating plan, in light of one's learning experience?

More perspicuously, the question is whether Conditionalization has the form

$$P_{old} \& \text{learn } E \Rightarrow \Box \Phi.$$

⁴It may be worth noting that even the construal of Conditionalization as a synchronic constraint on rational updating plans is not obviously compatible with time-slice rationality (Hedden 2013; Moss 2015). Very briefly, this is because time-slice rationality is committed to Impartiality in addition to Synchronicity, which is the principle that *from the agent's perspective*, it should not matter whether a belief or credence function is held by her future/past self or someone third party. But a constraint on updating plans does make that distinction. In particular, it says an agent's current credence should constrain or cohere with what credence function she thinks she will possibly adopt in the future.

or

$$\Box(P_{old} \& \text{learn } E \Rightarrow \Phi.)$$

where Φ is a claim about posteriors or updating plans.

3. *Evidential Certainty*. Is it an assumption of the principle that the agent becomes certain of what they learn, or a consequence thereof?⁵
4. *Factivity*. Does the principle assume that learning is factive (i.e. the information that one is asked to conditionalize on is true)?
5. *Partitionality*. Does the principle make assumptions about what one could have learnt otherwise? In particular, is it assumed that what one will possibly learn forms a partition of her doxastic space?

The answers to these questions are logically independent of each other. So there are 32 combinatorial possibilities in total (I give a partial list of some of them below). It is perhaps unsurprising that people who have discussed Conditionalization in the literature have not always had the same principle in mind. For instance, consider Ramsey (1926):

if p is the fact observed, my degree of belief in q after the observation should be equal to my degree of belief in q given p before, or by the multiplication law to the quotient of my degree of belief in $p \& q$ by my degree of belief in p .

The language suggests that Ramsey had in mind a diachronic, narrow-scope version of the rule that assumes factivity but not Evidential Certainty or partitionality. In contrast, Earman (1992) defines Conditionalization as the rule that

if it is learned for sure that E and if E is the strongest such proposition, then the probability functions P_{old} and P_{new} representing respectively degree of belief prior to and after acquisition of new knowledge, are related by $P_{old} = P_{new}(\cdot | E)$.

⁵Admittedly, “Evidential Certainty” is probably not the best terminological choice, given that what counts as evidence is itself a live controversy in epistemology. The bigger question lurking in the background is how we should model learning. There are at least two possible options:

- (a) The two-stage model (Howson and Urbach 2006; Schoenfield 2017), according to which an agent learns a piece of *exogenous* information (Schoenfield describes it metaphorically as “the world ‘flinging’ some information at the agent”), and processes that information endogenously.
- (b) The constraint model (Fraassen 1989; Joyce 1999; Gaifman and Vasudevan 2012), according to which a learning experience imposes some constraint on the agent’s posterior credences, and rational updating consists in selecting a single posterior credence function compatible with that constraint.

Conceptually, these two models differ with respect to what they take to be the *input* of rational learning: the two-stage model takes the input to be a piece of exogenous information (e.g. “roses are red”), whereas the constraint model takes the input to be a fact about the agent’s psychology (e.g. “the agent will assign credence one to the proposition “roses are red”). As transpired from this example, if we assume the constraint model in the background, then it would be natural to build Evidential Certainty into Conditionalization; the principle would be understood as a constraint on how rational agents should *propagate new certainties*. On the other hand, if we assume the two-stage model of learning, then it would seem less natural to assume Evidential Certainty (conceivably an agent may fail to be certain in the information the world flings at her).

Unlike Ramsey, Earman appears to take Conditionalization to be a constraint on *the relationship* between P_{old} and P_{new} . So his formulation is better understood as a wide-scope version of the principle. He also departs from Ramsey by building Evidential Certainty into the antecedent of the principle (it is assume that E is learned *for sure*), instead of factivity.⁶

Lastly, Lewis (1999) gave the following definition of Conditionalization (notation adapted):

Suppose that at time 0, you have a coherent belief function C . Let $E_1, \dots, E_n$ be mutually exclusive and jointly exhaustive propositions that specify, in full detail, all the alternative courses of experience you might undergo between time 0 and time 1. For each i from 1 to n , let C_i be the belief function you would have at time 1 if you had the experience specified by E_i , - that is, if E_i were the true one of $E_1, \dots, E_n$. You would *conditionalize* if, for any proposition X (in the domain of C),

$$C_i(X) = C(X|E_i) = \frac{C(XE_i)}{C(E_i)}.$$

So unlike both Ramsey and Earman, Lewis assumes that the conditionalizer will learn a true member of a partition. He also seems to take the norm to be first and foremost a norm on rational updating plans: it imposes a constraint on how you should update in response to each and every alternative courses of experience that you might undergo between time 0 and 1.

	D or P	W or N	Factivity?	EC?	Partitionality?
Ramsey (1926)	D	N	✓	✗	✗
Earman (1992)	D	W	✗	✓	✗
Lewis (1999)	P	N	✓	✗	✓

Some possible combinations:

- *Diachronic, wide-scope, factive, with Evidential Certainty and Partitionality built-in.*

Rationality requires that: if you have prior c at t_0 and between t_0 and t_1 you will learn with certainty which member of a partition $\mathcal{E}$ is true and nothing else, then, if you learn $E \in \mathcal{E}$, your credence at t_1 should be $c(\cdot|E)$.

- *Diachronic, narrow-scope, factive, with Evidential Certainty and Partitionality built-in.*

⁶Admittedly, there is some ambiguity. Earman uses the language “if it is learned for sure that E .” While I take this to be compatible with the possibility that learning is non-factive, Rescorla (forthcoming[b]) thinks this formulation “strongly suggests” that Earman had in mind a factive version of the principle.

If you have prior c at t_0 and between t_0 and t_1 you will learn with certainty which member of a partition $\mathcal{E}$ is true and nothing else, then, if you learn $E \in \mathcal{E}$, it is a rational requirement that your credence at t_1 be $c(.|E)$.

- *Diachronic, narrow-scope, factive, Evidential Certainty but no Partitionality.*

If you have prior c at t_0 and between t_0 and t_1 you learn with certainty the true information E and nothing else, then it is a rational requirement that your credence at t_1 be $c(.|E)$.

- *Plan, wide-scope, factive, with Evidential Certainty and Partitionality built-in.*

Rationality requires that: if you have prior c at t_0 and between t_0 and t_1 you will learn with certainty which member of a partition $\mathcal{E}$ is true and nothing else, then at t_0 you should plan to update on what you will learn by conditionalization (i.e. you plan to adopt credence $c(.|E)$ if you learn that E).

- *Plan, narrow-scope, factive, with Evidential Certainty and Partitionality built-in.*

If you have prior c at t_0 and between t_0 and t_1 you will learn with certainty which member of a partition $\mathcal{E}$ is true and nothing else, then at t_0 , it is a rational requirement that you plan to update on what you will learn by conditionalization.

- *Plan, narrow-scope, factive, Evidential Certainty but no Partitionality.*

If you have prior c at t_0 and between t_0 and t_1 , at each possible state w , you will learn with certainty the true information $E(w)$, then at t_0 , it is a rational requirement that you plan to update on what you will learn by conditionalization (i.e. you plan to adopt credence $c(.|E(w))$ at w).

Lastly, there are many arguments that purport to support Conditionalization—the Dutch Strategy Argument (Lewis 1999), the Pragmatic Expected Utility Argument (Brown 1976), the Expected Accuracy Argument (Greaves and Wallace 2006), the Accuracy Dominance Argument (Briggs and Pettigrew, forthcoming), the Symmetry Argument (Hughes and Fraassen 1984) and the Minimum information Argument (Diaconis and Zabell 1982). Which version (if any) of Conditionalization does each of these arguments support? The following table provides a quick summary:⁷

⁷One complication: although Lewis appears to state Conditionalization as a narrow-scoped norm, the diachronic Dutch Book argument only supports a wide-scoped conclusion: an agent who starts off with an incoherent prior would be similarly Dutch-bookable even if they update by conditionalization, and so Dutch-bookability only reveals defects in the *combination* of

	D or P	W or N	Factivity?	EC?	Partitionality?
Dutch Book (Lewis 1999)	P	W	✓	✗	✓
Dutch Book (Rescorla, forthcoming(a))	P	W	✗	✗	✓
Pragmatic Expected Utility	P	N	✓	✗	✓
Expected Accuracy	P	N	✓	✗	✓
Accuracy Dominance	P	W	✓	✗	✓
Symmetry	P	N	✗	✓	✗
Minimum Information	D	N	✗	✓	✗

the agent's prior and updating strategy.

Chapter 1

Conglomerability and Countable Additivity

The strength of our beliefs comes in degrees. I am very confident that I have hands, less so that there are aliens. Just as there seem to be coherence norms for outright beliefs – e.g. one should not believe a proposition and its negation – so too there seem to be coherence norms for degrees of belief (credences). *Prima facie*, it seems irrational to be highly confident that there are no aliens, and equally highly confident that there are aliens on Mars.¹

One widely accepted coherence norm for credences is that they should be finitely additive:

FA For any finite collection of disjoint propositions, an agent’s credence in their disjunction should equal the sum of her credences in each of the disjuncts.

Some theorists think FA is too weak. Instead, rational credences should be additive for any *countable* collection of disjoint propositions:

CA For any countable collection of disjoint propositions, an agent’s credences in their disjunction should equal the sum of her credences in each of the disjuncts.

There are at least two reasons why the truth of CA may be of philosophical importance. First, some philosophers think that human minds can grasp infinitely many propositions (Field 1978, p.16). If this is true, then a natural question to ask is whether one’s degrees of belief in those propositions are subject to coherence norms stronger than FA. The naive answer seems to be yes. It is an open question whether the

¹Some deny that there are coherence norms in general (e.g. Kolodny 2007, 2008) or coherence norms for credences in particular (e.g. Caie (2013)). Since I am focusing here on a debate where both parties accept that rational credences should be finitely additive, I shall set aside those worries for now.

universe is infinite. Suppose it is. Pre-theoretically, it would seem irrational to be certain (i.e. with credence 1) that there are aliens, and yet be certain that there are no aliens on planet P_i , for every individual P_i other than the Earth.

Second, CA plays an important role in scientific practices. Following Kolmogorov, most contemporary probability textbooks include countable additivity as an axiom of probability (e.g. Billingsley 2008; Cinlar 2011). Moreover, *unlike* Kolmogorov, who took countable additivity to be devoid of “empirical meaning” and assumed it for technical expedience (1956, p.14), his followers adopted a much more sober attitude towards the axiom. Indeed, Halmos considered countable additivity to be “a restriction without which modern probability theory could not function” (1970, 187), and he seems right. The principle is indispensable for proving various foundational results in modern probability theory and its applications, including the zero-one laws (cf. Cinlar 2011), the merging of Bayesian priors (Blackwell and Dubins 1962) and the existence of Nash equilibrium in a large economy (Khan et al. 2020).² Thus, Juhl and Kelly (1994) conclude, “countable additivity is actually a powerful axiom...that should be subjected to the greatest possible philosophical scrutiny.”

This and the following chapter take up that task. In this chapter, I subject to scrutiny one of the strongest arguments for CA – the conglomerability argument – and argue that it does not succeed. In Chapter 2, I put forward two new arguments against CA. Taken together, these two chapters put pressure on the idea that countable additivity should be upheld as a universal rational norm.

Before proceeding, a disclaimer: I am not a strict finitist. That is, I think in *some* cases, it can be a (defeasible) rational requirement to have degrees of belief that are countably additive. I am only skeptical of the claim that failures of countable additivity are *always* rationally criticizable. Admittedly, once we retreat from the universal claim, we face the question of where to draw the line. I think this is a hard question, but also a question for everyone. For fundamentally I think this is a question about the limits of infinitary idealization in our philosophical and scientific theorizing. On the one hand, few of us want to ban infinities from our scientific models and practices. On the other hand, we would like to stop somewhere, for infinities are weird, and some infinities are weirder than others. I hope my discussion will bring the “stopping point” problem into sharper focus.

²There have been attempts to prove some convergence and limiting results without the assumption of countable additivity, most notably in the strategic framework introduced in Dubins et al. (2014). See also Purves and Sudderth (1976), Chen (1976), and Purves and Sudderth (1983).

1 Introduction

One of the strongest arguments in favor of CA is the conglomerability argument.³ Roughly, the idea is that rational agents should have credences that are countably additive because otherwise they would be disposed to reason to a foregone conclusion (Kadane et al. 1996). One response to the conglomerability argument is that it begs the question. While a countably additive probability measure is conglomerable in any countable partition, it may not be conglomerable in a partition of higher cardinality (Seidenfeld et al. 2014). So proponents of CA must concede that inconglomerability is not always a rational defect – it is bad only when the relevant partition is countable.⁴ But what is so special about conglomerability in countable partitions, other than that it justifies CA?

There is one wrinkle with the question-begging objection: Seidenfeld et al.’s result presupposes the coherent theory of conditional probability (Finetti 1970; Dubins et al. 1975). Kolmogorov (1956) developed an alternative theory of conditional probability, which has also become the received orthodoxy among scientists and statisticians. One natural question to ask is whether a version of the question-begging objection would still arise, if one adopts this alternative theory of conditional probability. I argue it does. To this extent, I conclude that there is no path from conglomerability to countable additivity.

Here is the plan: §2 defines conglomerability and coherent probability. §3-4 recapitulate the conglomerability argument and the arbitrariness objection. §5-6 sketch Kolmogorov’s alternative theory of conditional probability, and argue it does not help salvaging the conglomerability argument. §7 concludes with general philosophical reflections.

2 Preliminaries

2.1 Conglomerability

Debby. Debby is solving a homicide. She has narrowed it down to two primary suspects: the father and the girlfriend. Moreover, she has just found a fingerprint at the crime scene. The forensic lab is now running through the database to find a match. Suppose Debby thinks to herself:

“Right now I am 50-50 between the father and the girlfriend. However, as soon as the lab gets back to me, no matter what the result is, I will be able to confirm that it is the father and not the girlfriend.”

³To my knowledge, the conglomerability argument was first introduced by Finetti (though de Finetti himself was unpersuaded by the argument). Proponents of the arguments include Easwaran (2013a), Titelbaum (2015), and Hoek (2020a)

⁴A predecessor of Seidenfeld et al. (2014) is Arntzenius et al. (2004) who proved that the uniform distribution over the unit square must be inconglomerable in some uncountable partitions under weak non-bias constraints.

Debby seems irrational. On the one hand, she thinks she can't rule out the girlfriend at the moment; on the other hand, she thinks she will be able to rule out the girlfriend once she hears back from the lab, irrespective of what they say. But if it does not matter what the lab finds out about the fingerprint, why can't she rule out the girlfriend now, before getting the test result?

We can dramatize this intuition by introducing a fictitious bookie. Suppose Debby is willing to bet on the fair price determined by her credences. Then before the lab result comes back, she is willing to sell a \$1 bet to the bookie that pays \$2 if the girlfriend is the murderer. After the result comes back, no matter what happens, she is willing to buy back the same bet at a higher price, and suffers a sure loss as a result.

Debby's credences are *inconglomerable*: there is a proposition (i.e. that the girlfriend is the murderer) and a partition of the logical space (i.e. what the lab might report back to her) such that her unconditional credence in that proposition is lower than her conditional credences in that same proposition given every single partition element.⁵ Our pre-theoretic assessment of Debby suggests that inconglomerability is a sign of irrationality. The question that we are going to investigate in this chapter is whether this intuitive judgment entails and/or coheres with CA.

To do so, I need to say more precisely what I mean by "conditional credence." Let Ω be a set of states. For the purposes of this chapter, I am going to define Ω stipulatively as the set of metaphysically possible states compatible with the agent's knowledge (in a fixed context). Propositions are represented as subsets of Ω . (Throughout this chapter, I shall abuse notation and use set-theoretic operations " $\cap$ " or " $\cup$ " and logical connectives " $\wedge$ " or " $\vee$ " interchangeably.)

Let $\mathcal{P}(\Omega)$ denote the powerset of Ω . A **conditional credence function over Ω** is a partial, binary, real-valued function

$$C(.|.) : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \setminus \{\emptyset\} \rightarrow \mathbb{R}.$$

with $C(A|B)$ representing the conditional credence that the agent assigns to A , given that B is true. In particular, for every $A \subset \Omega$, $C(A|\Omega)$ represents the agent's conditional credence in A given the tautology proposition (tautology with respect to what the agent knows). Following conventions of the literature, I shall call this value the agent's **unconditional credence in A** , and write it as " $C(A)$ " instead of " $C(A|\Omega)$."

Fix C and Ω . Let $\mathcal{F}_C$ denote the set of propositions with well-defined unconditional credences under C . In general, I do not require that $\mathcal{F}_C$ be equal to $\mathcal{P}(\Omega)$. An agent may not have precise degrees of beliefs in

⁵Throughout this chapter, I shall take conglomerability to be a synchronic constraint on the relation between one's conditional and unconditional credences at a time. There is a different interpretation, according to which conglomerability is a diachronic constraint the relation between one's prior at a time t and one's possible posteriors at a later time t' . This distinction would not make a difference if we assume Conditionalization, i.e. one's posterior credence after learning E and nothing else should equal one's prior conditional credence given E . However, there is a sense in which the synchronic conglomerability argument is stronger than the diachronic conglomerability argument. For an anti-CA theorist can in principle get around the diachronic conglomerability argument by rejecting Conditionalization (cf. Howson (2014)), whereas one cannot get around the synchronic conglomerability argument in this way.

some propositions in virtue of having never considered them before, even if they are logical combinations of possibilities that she has not ruled out on the basis of her knowledge. However, I am going to make the idealizing assumption that $\mathcal{F}_C$ always forms a sigma-algebra, i.e. it contains the tautology proposition Ω and is closed under negation and countable disjunction. Another assumption I shall make (which is again standard in the literature and won't affect the substance of this discussion) is that an agent has well-defined conditional credences given B , only if she has well-defined unconditional credence in B . In other words, C is only well-defined on $\mathcal{F}_C \times \mathcal{G}$ for some $\mathcal{G} \subset \mathcal{F}_C$.

Let $\mathcal{E} \subset \mathcal{F}_C$ be a partition of Ω . We say C is **conglomerable in $\mathcal{E}$** (or **$\mathcal{E}$ -conglomerable**) just in case

$$\forall A \in \mathcal{F}, \inf_{E \in \mathcal{E}} C(A|E) \leq C(A) \leq \sup_{E \in \mathcal{E}} C(A|E).$$

In English: an agent's credences are $\mathcal{E}$ -conglomerable just in case she is not disposed to change her mind in a predictable direction, irrespective of which element of $\mathcal{E}$ she learns to be true.

Having defined what conglomerability is, let me say a bit more about the notion of probabilistic coherence in this set-up.

2.2 Coherence

Recall our intuitive gloss of Probabilistic Coherence (cf. Introduction): rational agents should have degrees of belief that satisfy the standard (Kolmogorovian) axioms of probability. This formulation was fine, if we represent credal states by *unary* functions from propositions to real numbers (in other words if we take *unconditional* credences as primitive). But if we model credal states with binary functions (i.e. taking conditional credences as primitive), the axioms may seem too weak to capture the idea of coherence. In particular, it might appear that coherence of a conditional credence function C requires, not only that $C(\cdot|B)$ is coherent whenever it is well-defined, but also that any two well-defined $C(\cdot|B)$ and $C(\cdot|B')$ cohere with each other in the right way.

As we shall see, what counts as “probabilistic coherence” for conditional credences turns out to be controversial. In fact, there is a sense in which that question is equivalent to the question whether conglomerability is a rational requirement. But we are getting ahead of ourselves. For present purposes, let me adopt the following definition of coherence due to de Finetti):⁶

⁶More precisely, de Finetti defined coherence in terms of Dutch-bookability: a conditional credence function (modelled as a linear functional P that maps a set of (bounded) conditional bets $X|H$ to real numbers $P(X|H)$) is coherent just in case there are no n conditional bets $X_i|H_i$ and real numbers $c_1, \dots, c_n$ such that, given that one of H_i obtains, accepting $c_i(X_i|H_i)$ at price $c_i P(X_i|H_i)$ would guarantee the agent a sure loss. More precisely: there are no $X_1|H_1, \dots, X_n|H_n$ such that

$$\sup_{w \in H} \sum_{i=1}^n c_i 1_{H_i}(w) (X_i(w) - P(X_i|H_i)) < 0.$$

Definition 1. Let Ω be a sample space and C a conditional credence function on Ω . We say C is a **coherent conditional probability** if

- $\forall B \in \mathcal{F}_C \setminus \{\emptyset\}$, $C(.|B)$ is a well-defined finitely additive probability function with $C(B|B) = 1$;
- (Multiplicative Axiom) $\forall F, G, H \in \mathcal{F}_C$ with $H \neq \emptyset$ and $G \cap H \neq \emptyset$,

$$C(F \cap G|H) = C(F|G \cap H)C(G|H).$$

Note that, if C is regular, i.e. $C(A) > 0$ for all $A \in \mathcal{F}_C$ and $A \neq \emptyset$, then C is a coherent conditional probability just in case $C(.|\Omega)$ is a finitely additive probability function and $C(.|B)$ is given by the ratio formula. In other words, in the simple setting, Coherence is equivalent to the conjunction of the two Bayesian dogmas Probabilistic Coherence and Conditionalization.

2.3 Countable additivity

Let me conclude this section with a brief discussion of the formulation of CA in this framework. Initially, one might have thought that a precise statement of CA should look something like the following:

CA-COND A conditional credence function C is rational only if $C(.|B)$ is countably additive whenever it is well-defined.

However, in this chapter, I shall focus primarily on the following weaker principle:

CA A conditional credence function C is rational only if $C(.|\Omega)$ is countably additive.

CA says that rationality requires countable additivity of unconditional credences; CA-COND says that conditional credences need be countably additive as well. The two principles are equivalent if we assume Conditionalization (i.e. rational agents should revise their credences by conditionalization) together with the meta-normative principle that I call Closure: if one starts out with a rational prior and updates rationally, then one is guaranteed to end up with a rational posterior.

I think Closure is compelling (cf. Chapter 2). So a Bayesian should endorse CA-COND if she accepts CA. However, CA-COND also turns out to be a very strong principle. In particular, under very weak assumptions, the principle is incompatible with the rational permissibility of being indifferent given finitely many possibilities, or that of assigning full credence to what one learns. I shall discuss these negative arguments against CA-COND in the next chapter, and restrict my attention to CA here in order to avoid

where $H = \bigcup_i H_i$. It turns out that, when P takes as its domain $\{A|B : A \subset \Omega, B \neq \emptyset\}$, then P is coherent in the above sense just in case it is coherent in the sense of Definition 1 (cf. Berti et al. 2017).

those complications. Clearly, if there are no conglomerability arguments for CA, then *a fortiori* there are no conglomerability arguments for CA-COND.

3 The conglomerability argument

Let's start with the simple case. Fix a conditional credence function C over a sample space Ω . Let $\mathcal{E} \subset \mathcal{F}_C$ be a finite partition of Ω . Suppose $C(E) > 0$ for all $E \in \mathcal{E}$. In this case, it is easy to check that C is $\mathcal{E}$ -conglomerable if it is a coherent conditional probability.⁷ In other words, the following is a consequence of Coherence:

Finite Conglomerability. C is rational only if C is conglomerable in any finite partition $\mathcal{E} \subset \mathcal{F}_C$.

Intuitively, whether inconglomerability is a rational defect should not depend on the cardinality of the relevant partition. Suppose there are infinitely many possible individuals whose fingerprint could potentially match the one found at the crime scene. Still, Debby's beliefs would strike us as irrational and for the same reason: If she will be able to rule out the girlfriend irrespective of which one of the (infinitely many) possible matching outcomes turn out to be true, then that means the test result does not matter in the first place; she should be able to rule out the girlfriend now, before hearing back from the lab.

The above consideration suggests that, if Finite Conglomerability is true, then the following should be a rational norm as well:

Countable Conglomerability. C is rational only if C is conglomerable in every countable partition $\mathcal{E} \subset \mathcal{F}_C$.

However, Countable Conglomerability is false, if it is rationally permissible to have credences that are not countably additive. To see why, consider the following case:

Lotti. Lotti is solving a different mystery: she is trying to figure out how a fair coin landed the last time when tossed. The only information she has is that the coin was used by Dr. Ubins to decide how he was to go about selecting a natural number. She also knows his protocol: Heads means the number was drawn randomly; Tails means the doctor tossed another fair coin, and picked the number n , where n was the first time the coin landed heads. Fortunately, someone found the number picked by Dr. Ubins, and is going to message Lotti shortly.

⁷For completeness: if C is a finitely additive probability function and $\mathcal{E}$ is finite, then

$$C(A) = \sum_{E \in \mathcal{E}} C(A \& E) = \sum_{E \in \mathcal{E}, C(E) > 0} C(A|E)C(E).$$

It follows that $C(A)$ must be bounded below by $\inf_{E: C(E) > 0} C(A|E)$ and above by $\sup_{E: C(E) > 0} C(A|E)$, which is a smaller interval than $[\inf_E C(A|E), \sup_E C(A|E)]$. Note that this argument goes through irrespective of how one defines $C(\cdot|E)$ when $C(E) = 0$.

Lotti thinks to herself: “Right now I am 50-50 about whether the coin landed heads or tails – after all, I know it is a fair coin. However, as soon as I get the message, I can be sure that the coin landed tails.”

At first glance, Lotti seems as irrational as Debby. However, the anti-CA theorist must deny this symmetry. For one principal motivation for her position is precisely that, in cases like *Lotti*, it is rationally permissible for Lotti to be indifferent about what number was picked by Dr. Ubins, if the coin landed heads, i.e. it is permissible for her conditional credence given Heads to be a uniform distribution over the natural numbers. But then, assuming Conditionalization, it follows that it is rationally permissible for Lotti to be fully confident that the coin landed Tails, no matter what the message says:

$$C(H|N = n) = \frac{C(H, N = n)}{C(H, N = n) + C(T, N = n)} = \frac{0}{0 + \frac{1}{2^n}} = 0.^8$$

So the anti-CA theorist is committed to drawing a normative distinction between Debby and Lotti: one is rationally criticizable, the other one is rationally unimpeachable. (The CA theorist avoids this embarrassment by denying that Lotti’s situation is coherent in the first place. For if CA is true, then it is never rationally permissible to assign uniform zero credence over a countable sample space. Assuming Principal Principle (i.e. chance guides rational credence), it follows that it is impossible to select a natural number randomly such that the chance of each number being selected is uniformly zero.)

Lotti shows that *some* failures of countable additivity lead to inconglomerability. As it turns out, all failures of countable additivity entail inconglomerability in some countable partition (Schervish et al. 1984; Hill and Lane 1985; Zame 1988). This result motivates the following argument for CA:

P1 Finite Conglomerability is true.

P2 If Finite Conglomerability is true, then Countable Conglomerability is true.

P3 Countable Conglomerability is true only if CA is true.

∴ CA is true.

4 Begging the question

An immediate worry about the conglomerability argument is that, if P2 were true, then the following stronger claim also seems plausible:

⁸In fact we can dramatize the example even more: Note that the same conclusion holds irrespective of the initial bias of the coin. As long as Lotti thinks it is possible for the coin to have landed tails, she will be confident that the coin landed tails after receiving that message, *even if* she were *a priori* confident that the coin landed heads, or that it is 0.999999 biased towards heads.

P2* If Finite Conglomerability is true, then Total Conglomerability is true.

However, it turns out that the CA theorist can't accept P2*, for the exact same reason that her opponent can't accept P2. This is because Seidenfeld et al. (2014) have shown that, under certain structural assumptions, for any cardinal κ , a probability measure is conglomerable in any κ -partition only if it is κ -additive. In particular, under those assumptions, a real-valued probability measure is totally conglomerable only if it is atomic – it assigns probability 1 to countably many atomic possibilities. Almost all probability distributions used by scientists and statisticians are nonatomic, Lebesgue measure being one of the most prominent examples. So P1 and P2* jointly entail that scientists are more than often irrational. But, as noted at the beginning of this chapter, one principal argument for CA is by abduction: It features prominently in our scientific practices, and so the burden of proof is on its opponent to show why it is nevertheless not true. So it would appear that the CA theorist can't accept P2*, on pain of self-undermining.

To give the reader a more concrete sense of the failure of Total Conglomerability, let me work through Seidenfeld et al.'s result with respect to an example:

Corinne. Corinne is studying the precise bias of a coin, about which she knows nothing. Let $\Omega = [0, 1]$, where each $x \in \Omega$ represents the possibility that the bias of the coin is x . Let C be Corinne's prior credence function over Ω . Suppose C is mildly smooth in the following two senses:

1. $C(\cdot|\Omega)$ is the uniform Lebesgue measure;
2. for any two distinct values $x, y \in [0, 1]$, given that the true bias of the coin is x or y , Corinne is not fully confident that it is x rather than y or *vice versa*, i.e. for every $x, y \in [0, 1]$, $C(x|x \vee y) \neq 1 \neq C(y|x \vee y)$.⁹

Let H be the proposition that the coin is biased towards tails. It turns out that, for any $\epsilon > 0$, there is a function f that maps every $x > \frac{1}{2}$ to a disjoint finite subset $f(x) \subset [0, 1/2]$ (i.e. $x \neq y \Rightarrow f(x) \cap f(y) = \emptyset$), such that $C(x|f(x) \cup \{x\}) < \epsilon$. It follows that C is inconglomerable in the following partition:

$$\mathcal{E} = \{f(x) \cup \{x\} : x \in H\} \cup \{\{y\} : y \notin H\} \cup \bigcup_{x \in H} f(x).$$

In particular, $C(H) = \frac{1}{2}$ but for every $E \in \mathcal{E}$, $C(H|E) = 0$ or ϵ .

To bring the point home, imagine a reliable informant who knows which proposition in $\mathcal{E}$ is true, i.e. if the coin has a bias $x \geq 1/2$ and not in the union of the image of f , then they know what it is; otherwise, they

⁹In other words, C is weakly regular in the sense of Builes (forthcoming)). Weak regularity entails that C satisfies one of the sufficient conditions in Seidenfeld et al. (2014)

know that it belongs to the finite list $\{x\} \cup f(x)$. The informant is to message the true answer to Corinne. It follows that, before she receives the message, Corinne would think to herself,

“Right now I am 50-50 about whether the coin is biased towards heads or tails. However, as soon as I get the message, no matter what I learn, I will be certain or confirm with very high confidence $(1 - \epsilon)$ that the coin is biased towards heads.”¹⁰

Corinne seems as irrational as Debby and Lotti. Admittedly, the set of propositions that would cause Corinne to have the seemingly irrational set of beliefs is much larger. But it is unclear why that feature of her situation should carry any normative weight. Moreover, even if it does, it would appear that the same consideration can be used to break the symmetry between Debby and Lotti. Why should inconglomerability in countable partitions be a rational defect, if inconglomerability in uncountable partitions is rationally unimpeachable?

So proponents of the conglomerability argument face a dilemma. Either they accept that Corinne is irrational, or they don't. If they do, then they must deny the rational permissibility of Corinne's prior, as well as other intuitively plausible alternatives. In fact, the only options left are almost exclusively those that have their probabilities concentrated on countably many possibilities. Extreme *a priori* bias is not only rationally permitted, but rationally mandated. As noted before, this appears to be an unpalatable consequence, especially for proponents of CA.

On the other hand, suppose proponents of the argument insist that Corinne is rational, despite violating Total Conglomerability. It follows that, like their opponents, they too think that inconglomerability is not a universal rational defect; rationality requires conglomerability in some partitions, but not others. But if we have to draw the line somewhere, why draw it between Lotti and Corinne, rather than Debby and Lotti? What makes Countable Conglomerability a special stopping point, other than the fact that it entails CA?

5 Question-sensitivity

Let's take a step back. One assumption we made at the beginning is that the credal state of a rational agent can be represented by a (partial) binary function $C(\cdot|\cdot)$ that assigns real numbers to pairs of propositions.

One could challenge this assumption, based on the following consideration.¹¹

¹⁰In fact, Kadane et al.'s result entails a stronger claim: assuming the theory of coherent conditional probability, in *Corinne*, Total Conglomerability is incompatible with the rational permissibility of having a prior C that is weakly regular over an uncountable set of positive probability, i.e. if C is totally conglomerable, then for every uncountable S , if $C(S) > 0$, then there exists $x, y \in S$ such that $C(x|x \vee y) = 1$ or $C(y|x \vee y) = 1$. In particular, it follows that, given Total Conglomerability, the set of points whose comparative probabilities are non-extreme must receive zero unconditional probability.

¹¹The example is adapted from Myrvold (2015)

Factory. A factory produces perfect cubes. Suppose one is informed by a reliable source that the face area of a randomly chosen cube follows a uniform distribution between 0cm^2 and 1cm^2 (exclusive). A cube is picked at random. What is the probability that its face area is $\frac{1}{4}\text{cm}^2$, given that it is $\frac{1}{4}\text{cm}^2$ or $\frac{3}{4}\text{cm}^2$?

Naively, one might have thought that the answer is $\frac{1}{2}$, i.e.

$$P(A = \frac{1}{4} | A = \frac{1}{4} \vee A = \frac{3}{4}) = \frac{1}{2}$$

Moreover, there seems to be a *prima facie* justification for this intuition. Pick any $\epsilon > 0$. The probability that the face area of the cube is $(x \pm \epsilon)\text{cm}^2$ is positive for any x . So we can compute the following conditional probability using the ratio formula:

$$P(A = \frac{1}{4} \pm \epsilon | A = \frac{1}{4} \pm \epsilon \vee A = \frac{3}{4} \pm \epsilon) = \frac{P(A = \frac{1}{4} \pm \epsilon)}{P(A = \frac{1}{4} \pm \epsilon) + P(A = \frac{3}{4} \pm \epsilon)} = \frac{\epsilon}{2\epsilon} = \frac{1}{2}.$$

But intuitively, $P(A = \frac{1}{4} | A = \frac{1}{4} \vee A = \frac{3}{4})$ is just a limiting case of $P(A = \frac{1}{4} \pm \epsilon | A = \frac{1}{4} \pm \epsilon \vee A = \frac{3}{4} \pm \epsilon)$. So its value should be $\frac{1}{2}$.

Now scratch that information. Suppose one is told instead that the side length of a randomly chosen cube follows a non-uniform distribution with density $f(x) = 2x$. What is the probability that a cube picked at random has a side length of $\frac{1}{2}\text{cm}$, given that it has a side length of $\frac{1}{2}\text{cm}$ or $\frac{\sqrt{3}}{2}\text{cm}$?

This time, one might have thought that the answer should be lower than $1/2$, and this intuition is again vindicated by the previous method: for any small $\epsilon > 0$, the probability that the side length of the cube is $(\frac{\sqrt{3}}{2} \pm \epsilon)\text{cm}$ is positive and greater than the probability that the side length is $(\frac{1}{2} \pm \epsilon)\text{cm}$. Therefore,

$$P(L = \frac{1}{2} \pm \epsilon | L = \frac{1}{2} \pm \epsilon \vee L = \frac{\sqrt{3}}{2} \pm \epsilon) < \frac{1}{2}.$$

So, if we take

$$P(L = \frac{1}{2} | L = \frac{1}{2} \vee L = \frac{\sqrt{3}}{2})$$

to be a limiting case of

$$P(L = \frac{1}{2} \pm \epsilon | L = \frac{1}{2} \pm \epsilon \vee L = \frac{\sqrt{3}}{2} \pm \epsilon),$$

we arrive at the conclusion that its value should be smaller than $1/2$.

However, a cube has face area $\frac{1}{4}\text{cm}^2$ ($\frac{3}{4}\text{cm}^2$) if and only if it has side length $\frac{1}{2}\text{cm}$ ($\frac{\sqrt{3}}{2}\text{cm}$). So these two judgments are incompatible with each other. One must give, though it is not clear which one (see Chapter 3 for further discussion).

Examples of this kind – broadly known as *the Borel-Kolmogorov paradox* – have moved some theorists to deny that the conditional probability of A given B is always well-defined. Most notably, Kolmogorov himself concludes from this paradox that “the concept of a conditional probability with regard to an isolated given hypothesis whose probability equals zero is inadmissible.” Instead, the value of $P(A|B)$ depends on the choice of the conditioning partition to which B belongs as a member (or, in the more general setting, the conditioning sigma-algebra that contains B as an atom). Different choices of partitions may yield different values of $P(A|B)$; there is no “the” conditional probability of A given B , independently of the choice of a partition.

Kolmogorov was a frequentist (i.e. he thought of probabilities as objective features of events, identifiable with (hypothetical) limiting frequencies). But one could give his proposal an epistemic interpretation. If we identify conditional probabilities with rational conditional credences, then one way of reading Kolmogorov’s suggestion is that rational conditional credences should be sensitive to the choice of a partition. For present purposes, we can interpret a partition as a question (cf. Chapter 3 for other interpretations). Each partition element corresponds to a possible answer of the question. For instance, the question “What is the precise bias of a coin?” may be represented by the partition $\{\{x\} : x \in [0, 1]\}$ of the unit interval, where each x represents the possibility that the coin has a bias of exactly x .

The rough idea of the Kolmogorovian proposal is this: One can’t conditionalize on B *simpliciter* – not at least when B receives zero unconditional credence. Rather, one *must* conditionalize on B *qua* the true and complete answer to some background question Q . Moreover, the choice of Q can be epistemically relevant. There can be cases – *Factory* being one example – where it is rational to be 0.5 confident in A given B , when B is viewed as an answer to Q , but less than 0.5 confident in A given B , when B is taken to be an answer to a different question Q' .

Before delving into the formal details of the Kolmogorovian proposal, let me pause here to contrast it with a similar idea that has been proposed in the literature. Some have argued that unconditional rational beliefs/credences can be question-sensitive (Yalcin 2016; Hoek 2020b; Elga and Rayo, forthcoming). For instance, when asked “How to spell ‘dreamt’?,” one may be able to answer in a heartbeat that it ends in “-mt.” But when asked “Which words end in ‘-mt’?,” one may stumble and not have a clue. One intuitive explanation of this phenomenon is that the person believes that “dreamt” ends in “-mt” in response to the first question, but fails to have (or access) that belief in response to the second question. Moreover, one might think this person is not rationally criticizable. (If one does not share that intuition, consider some mathematician who is uncertain whether 1234567 is a prime, but when asked to compute 127×9721 , is able to do the mental calculation and form the belief that it is 1234567.) There are different ways of vindicating this judgment. Some argue that rational assessment is always sensitive to the context of inquiry (Hoek

2020b); others think the agent in this example is fragmented, and fragmentation is not an intrinsic rational defect (Yalcin, forthcoming).¹² For the lack of a better terminology, I shall group these views under the heading “fragmentationalism.”

For what it’s worth, I am attracted to fragmentationalism, though I won’t defend it here (or anywhere in this dissertation). Instead, let me note two important respects in which fragmentationalism differs from the Kolmogorovian proposal. First, fragmentationalism is about rational permissibility; rational agents *may* have different beliefs towards the same proposition in response to different questions. By contrast, the Kolmogorovian proposal stated above says that, in some cases, rational agents *must* adopt different conditional credences given the same proposition, depending on which question it addresses. So the Kolmogorovian proposal is much stronger, and more intriguing to that extent. For, pre-theoretically, it would appear that even if one may believe differently in response to different questions, one can’t possibly be criticized for having the same belief – conditional or not – in response to different questions. Coherence may not be a rational requirement, but it can’t be a rational sin either.

Second, fragmentationalism often goes with a caveat: Rational agents may have different beliefs towards the same proposition in response to different questions, *provided* that these questions are not all salient or foregrounded in the context. The fragmentationalist usually concedes that, if one is asked which word ends in “-mt” immediately after one spells “dreamt” correctly, then it would be irrational for one not to believe that “dreamt” is an answer. More generally, one main motivation for relativizing rational beliefs to questions is to accommodate rational logical ignorance. The thought is, rational agents may fail to believe logical consequences of their beliefs, in virtue of not having been exposed to the relevant question, or not having the question made salient to them in the relevant context. However, once the questions are at the forefront of their mind, there does seem to be some rational pressure to have their attitudes towards those questions be consistent with each other. (In fact, one might think such coherence requirement is a necessary condition for rational inquiry – an agent cannot make progress in an investigation, unless how she answers one question rationally constrains how she proceeds thereafter.)

By contrast, the Kolmogorovian proposal says that rational agents are sometimes required to adopt different conditional credences in A given B , depending on whether B is viewed as an answer to Q or Q' , *even when* both questions are salient to the agent. For instance, in *Factory*, the proposal says that whether one should be more confident that the face area of the cube is $3/4\text{cm}^2$, given that it is either $1/4\text{cm}^2$ or $3/4\text{cm}^2$, depends on whether one takes the latter to answer the question (i) “What is the face area of the cube?” or (ii) “What is the side length of the cube?”

¹²It may be worth noting there is another sense in which belief (but not credence) can be question-sensitive: given that one is x confident about a proposition p , whether one believes/should believe p may be question-sensitive (Yalcin 2010; Lin and Kelly 2012; Holguin 2020). Since I am focusing here on credences, I shall set this view aside for the moment.

(Formal caveat: in fact, these two two questions correspond to the same partition. This is a case where I think the formal representation of questions as partitions may be inadequate. Instead, in the infinite context, I propose to represent some questions as *limits* of sequences of partitions, which can themselves be thought of questions whose answers approximate the desired answer with some finite degree of precision. For example, in *Factory*, the two relevant sequences are: “What is the face area of the cube within a margin of error of ϵcm^2 ?” for every $\epsilon > 0$, and “What is the side length of the cube within a margin of error of ϵcm ?” for every $\epsilon > 0$.)

Let me concede right off the bat that the Kolmogorovian proposal is not supposed to be intuitive. Pre-theoretically, it would appear that how likely A is given B has nothing to do with whether B is an answer to Q or Q' . To dramatize this intuition: Suppose one asks a factory worker what is the face area of the cube. The worker replies that it is either $1/4\text{cm}^2$ or $3/4\text{cm}^2$. In response, one becomes $1/2$ confident that the face area of the cube is $1/4\text{cm}^2$. Suppose then the worker apologizes, “Sorry, I made a mistake – that was not an answer to you but to someone else, who was inquiring about the side length of the cube.” Intuitively, you shouldn’t change your credence as a result. The worker did not retract his assertion. He still stands by the claim that the face area of the cube is either $1/4\text{cm}^2$ or $3/4\text{cm}^2$. The fact that it was not supposed to address your question seems irrelevant. However, the Kolmogorovian seems to think that it is relevant – it should make you less confident that the cube has a face area of $1/4\text{cm}^2$.

I shall defer a fuller discussion of the Kolmogorovian proposal to Chapter 3. The most important point for our current purposes is that the Kolmogorovian thinks Corinne can adopt a rational prior that satisfies conditions 1 and 2 in the example, *and* is conglomerable over the partition

$$\mathcal{E} = \{f(x) \cup \{x\} : x \in H\} \cup \{\{y\} : y \notin H \bigcup_{x \in H} f(x)\}.$$
¹³

In particular, the proof that conditions 1 and 2 entail inconglomerability in $\mathcal{E}$ relies on the assumption that Corinne’s prior satisfies the Multiplicative Axiom. As we shall see, the Kolmogorovian endorses a coherence axiom that is strictly weaker than the Multiplicative Axiom. which in turn blocks the derivation.

6 Credal states

Let’s grant that rational conditional credence should be relativized to questions. To see how this assumption affects the conglomerability argument, we need to first incorporate this assumption into our formal model of credal attitudes. The proposal I am going to explore, which is largely in the spirit of Kolmogorov’s own

¹³Recall that f is a function that maps each $x \geq 1/2$ to a finite subset $[0, 1/2]$ with $f(x) \cap f(y) = \emptyset$ if $x \neq y$.

preferred theory of conditional probability (known as the theory of regular conditional distribution), is the following:

A **credal state** is represented by a partial function C , which maps partitions to conditional credence functions that agree given the tautology. More precisely: fix a sample space Ω . A credal state is a partial function

$$C : \mathcal{E} \mapsto C_{\mathcal{E}}(\cdot|\cdot) : \mathcal{P}(\Omega) \times \mathcal{E} \cup \{\Omega\} \rightarrow \mathbb{R}$$

with $C_{\mathcal{E}}(\cdot|\Omega) = C_{\mathcal{E}'}(\cdot|\Omega)$ whenever the two are well-defined.¹⁴

Intuitively, this last condition means that unconditional credences are not sensitive to questions.¹⁵ In particular, let $\mathcal{E}_C$ be the set of partitions on which C is well-defined. Then it follows from this assumption that $\mathcal{F}_{C_{\mathcal{E}}} = \mathcal{F}_{C_{\mathcal{E}'}}$ for any $\mathcal{E}, \mathcal{E}' \in \mathcal{E}_C$. I'll abuse notations and let “ $\mathcal{F}_C$ ” denote this sigma-algebra and “ $C(\cdot|\Omega)$ ” the unconditional credence function over $\mathcal{F}_C$ according to C . Lastly, as an idealizing assumption, we assume that $\mathcal{E}_C$ includes every partition that is a subset of $\mathcal{F}_C$, i.e. the agent has well-defined conditional credences relative to a question, if she assigns well-defined unconditional credences to each answer of that question.

Once we relativize conditional credences to partitions, we also need to modify the definitions of coherence and conglomerability accordingly. Here is a natural thought: Say a credal state C is **coherent** if, for every $\mathcal{E}$, $C_{\mathcal{E}}$ is coherent in the sense of Definition 1, i.e.

- $\forall E \in \mathcal{E} \setminus \{\emptyset\}$, $C_{\mathcal{E}}(E|E)$ is a well-defined finitely additive probability function with $C_{\mathcal{E}}(E|E) = 1$;
- (Local Multiplicative Axiom) For all $A \in \mathcal{F}_C$ and $E \in \mathcal{E}$, $C_{\mathcal{E}}(A \cap E|\Omega) = C_{\mathcal{E}}(A|E)C_{\mathcal{E}}(E|\Omega)$. In particular, if $C(B) = C_{\mathcal{E}}(B) > 0$, then $C_{\mathcal{E}}(A|B) = C(A \cap B)/C(B)$.

Clearly every coherent conditional credence function C can be thought of as a coherent credal state represented by the constant function

$$\tilde{C} : \mathcal{E} \mapsto C(\cdot|\cdot)$$

On the other hand, not every coherent credal state corresponds to a coherent conditional credence function. For a concrete illustration, let $\Omega = \{a, b, c, d\}$, $\mathcal{E}_1 = \{\{a, d\}, \{b, c\}\}$ and $\mathcal{E}_2 = \{\{a\}, \{b, c\}, \{d\}\}$. Suppose C

¹⁴In the more general case, where questions are represented by partitions as well as limits of partitions, a credal state is represented by a partial function

$$C : \mathcal{P}(\Omega) \times \mathcal{F} \mapsto C_{\mathcal{F}}(\cdot|\cdot) : \mathcal{P}(\Omega) \times \Omega \rightarrow \mathbb{R}$$

where $\mathcal{F}$ is a sub-sigma-algebra of $\mathcal{P}(\Omega)$ and for each w , $C_{\mathcal{F}}(\cdot|w)$ represents the agent's conditional credence given the information at w filtered through $\mathcal{F}$ (Rescorla 2018).

¹⁵We can of course relax this assumption. Maybe the more general framework will have C be a binary function from pairs of partitions to conditional credence functions. I shall leave this possibility for future investigation.

satisfies the following three equations:

$$C(a|\Omega) = 1, \quad C_{\mathcal{E}_1}(b|b \vee c) = 1, \quad C_{\mathcal{E}_2}(c|b \vee c) = 1$$

One can check that C can be extended to a function that represents a coherent credal state; but it cannot be extended to a coherent conditional credence function, since $C_{\mathcal{E}_1}$ and $C_{\mathcal{E}_2}$ disagree when conditionalizing on the same proposition, namely $b \vee c$.

A natural question one may ask at this point is when coherent credal states are coherent conditional credence functions. It turns out that the following condition is necessary and sufficient:

Global Coherence. For any $\mathcal{E}, \mathcal{E}' \subset \mathcal{F}_C$ and any $E \in \mathcal{E}, E' \in \mathcal{E}'$, if $E' \subset E$, then for any $A \in \mathcal{F}_C$,

$$C_{\mathcal{E}}(A \cap E' | E) = C_{\mathcal{E}'}(A | E') C_{\mathcal{E}}(E' | E).$$

In particular, for any $\mathcal{E}, \mathcal{E}'$ and $E \in \mathcal{E} \cap \mathcal{E}'$,

$$C_{\mathcal{E}}(.|E) = C_{\mathcal{E}'}(.|E).$$

Fact 1. Fix a coherent credal state C . C is globally coherent if and only if there exists a coherent conditional credence function $\tilde{C} : \mathcal{F}_C \times \mathcal{F}_C \setminus \{\emptyset\} \rightarrow \mathbb{R}$ such that for all $A \in \mathcal{F}_C$, partition $\mathcal{E} \subset \mathcal{F}_C$ and non-empty $B \in \mathcal{E}$,

$$\tilde{C}(A|B) = C_{\mathcal{E}}(A|B).^{16}$$

In other words, the Kolmogorovian framework allows us to factorize the coherence condition in Definition 1 into a conjunction of two different coherence requirements: the “local” requirement that says one’s conditional credence relativized to a partition should be coherent, and the global requirement that says one’s conditional credences given different partitions should cohere with each other in accord with Global Coherence. The Kolmogorovian accepts the requirement of local coherence, but rejects Global Coherence as a (universal) rational norm.

So much about coherence. Let’s turn now to countable additivity and conglomerability, about which I shall be brief. Let C be a credal state. We say C is **countably additive** if $C(.|\Omega)$ is countably additive; for any $\mathcal{E} \in \mathcal{E}_C$, C is **$\mathcal{E}$ -conglomerable** if $C_{\mathcal{E}}$ is $\mathcal{E}$ -conglomerable. It is straightforward to verify that a credal state C is conglomerable in any finite partition $\mathcal{E} \in \mathcal{E}_C$ if it is coherent; C is conglomerable in any countable partition $\mathcal{E} \in \mathcal{E}_C$ if it is coherent and countably additive. So finite additivity is sufficient for finite

¹⁶See Appendix A for the proof.

conglomerability, and countable additivity is sufficient for countable conglomerability.

6.1 Conglomerability for credal states

Now recall the original conglomerability argument:

P1 Finite Conglomerability is true.

P2 If Finite Conglomerability is true, then Countable Conglomerability is true.

P3 Countable Conglomerability is true only if CA is true.

$\therefore$ CA is true.

The observation we just made above shows that P1 is true in the Kolmogorovian framework. The pressing question now is whether P3 is true, i.e. whether countable additivity is necessary for countable conglomerability. If the answer is no, then the argument would be unsound, whether or not P2 is true.

Unfortunately, to my knowledge, this is an open question. We do know that, if $C(.|\Omega)$ is not countably additive and assumes infinitely many values, then there is a countably infinite partition $\mathcal{E}$ such that $C(E) > 0$ for all $E \in \mathcal{E}$ and $C_{\mathcal{E}}$ is not conglomerable in $\mathcal{E}$. The result also holds in full generality if we assume that C is globally coherent and in particular $C_{\mathcal{E}}(.|E) = C_{\mathcal{E}'}(.|E)$ for any $E \in \mathcal{E} \cap \mathcal{E}'$. But the CA theorist would presumably reject this last assumption. For Fact 1 shows that, given Global Coherence, the Kolmogorovian theory of conditional probability is equivalent to the coherent theory of conditional probability, and we would be back to square one.

To state the open problem more precisely (this part can be skipped): Let Ω be countably infinite. Let $\mathcal{U}$ be an ultrafilter over $\mathcal{P}(\Omega)$ that extends the countable-co-countable filter. Consider a credal state C such that $C(A|\Omega) = 1$ if $A \in \mathcal{U}$ and 0 otherwise. Then $C(.|\Omega)$ is finitely-valued and finitely but not countably additive. To show that countable additivity is necessary for countable conglomerability, we need to show in particular that there are no credal states C that satisfies the above constraints and is conglomerable over every countable partition of Ω . To my knowledge, it is an open question whether this is in fact true.¹⁷

So relativizing conditional credences to partitions may actually undermine rather than improve the tenability of the conglomerability argument; the argument may fail for mathematical reasons, even before we subject it to philosophical scrutiny.

There is however a different condition, closely related to conglomerability, that does entail countable additivity within the Kolmogorovian framework – countable disintegrability. The goal of the next section is

¹⁷It may be worth noting that any credal state that satisfies the above constraint would also be a “uniform” distribution over Ω , in the sense that $C(w) = 0$ for any $w \in \Omega$. So the CA theorist can’t dismiss this problem on the basis that credal states that satisfy the above constraints are “pathological” or intuitively irrational – doing so would again be begging the question against her opponent.

to explore whether the Kolmogorovian can find a path from disintegrability to countable additivity. I argue that the prospects are not bright. In particular, it turns out that, given other plausible assumptions, a CA theorist who accepts disintegrability principles must take conflicting stands on the truth of the Continuum Hypothesis.

6.2 Disintegrability for credal states

Recall that a credal state C is coherent just in case its relativization to every partition $\mathcal{E}$ is coherent in the sense of Definition 1, i.e. every $C_{\mathcal{E}}$ satisfies the following two requirements

- $\forall E \in \mathcal{E}$, $C_{\mathcal{E}}(\cdot|E)$ is a well-defined finitely additive probability function with $C_{\mathcal{E}}(E|E) = 1$;
- for all $A \in \mathcal{F}_C$ and $E \in \mathcal{E}$, $C_{\mathcal{E}}(A \cap E|\Omega) = C_{\mathcal{E}}(A|E)C_{\mathcal{E}}(E|\Omega)$.

But one might think this coherence condition is too weak. In particular, since the second requirement is always trivially satisfied when $C_{\mathcal{E}}(E|\Omega) = 0$, it imposes no constraints on $C_{\mathcal{E}}$ when $\mathcal{E}$ is an infinite partition whose elements receive zero unconditional credence under C .

Kolmogorov himself endorses a stronger constraint: for every $\mathcal{E} \subset \mathcal{F}_C$, $C_{\mathcal{E}}$ should satisfy **the integral formula**. In the current framework, this requirement is equivalent to the otherwise weaker condition known as **$\mathcal{E}$ -disintegrability**:

$$\forall A \in \mathcal{F}_C, C(A|\Omega) = \int_{\Omega} C_{\mathcal{E}}(A|\cdot) dC$$

where $C_{\mathcal{E}}(A|\cdot)$ is the real-valued function that maps each world $w \in \Omega$ to $C_{\mathcal{E}}(A|E)$, where E is the unique element in $\mathcal{E}$ that contains w . (Note that, in order for the integral on the right-hand side to be well-defined, $C_{\mathcal{E}}(A|\cdot)$ must be a measurable with respect to $\mathcal{F}_C$, meaning that for every rational r , the set $\{w : C_{\mathcal{E}}(A|\cdot) \leq r\}$ is a member of $\mathcal{F}_C$.) Assuming Conditionalization, this condition is in turn equivalent to the principle of Weak Reflection, according to which rational agents should have their priors equal their prior expectations of their posteriors.

It's not hard to see that $\mathcal{E}$ -disintegrability entails $\mathcal{E}$ -conglomerability. Indeed, it is strictly stronger: there are cases where a coherent $C_{\mathcal{E}}$ is $\mathcal{E}$ -conglomerable but not $\mathcal{E}$ -disintegrable. On the other hand, failures of disintegrability may seem just as bad from a rational point of view. Suppose Debby is 0.999 confident that the fingerprint found at the crime scene was not from the victim. As she waits for the result from the lab, she thinks to herself,

“Right now I am 50-50 between the father and the girlfriend. However, as soon as the lab gets back to me, if the fingerprint was not from the victim – and I am pretty sure that it was not –

then I will be able to confirm that it is the father and not the girlfriend. But if it was from the victim, then I shall remain indifferent between the two.”

In this case, Debby’s credences *are* conglomerable, though they are not disintegrable (in the partition induced by different possible lab reports). Nevertheless, her attitudes seems just as irrational and for the same reason: If she can rule out the girlfriend in all but a remote scenario, why can’t she do it now? What is the point of waiting for the lab result?

Moreover, while it is an open question whether countable additivity is necessary for countable conglomerability in the Kolmogorovian framework, it is straightforward to show that countable additivity is necessary for countable disintegrability. That is, for any credal state C , $C_{\mathcal{E}}$ is disintegrable in every countable partition $\mathcal{E} \subset \mathcal{F}_C$ only if $C(\cdot|\Omega)$ is countably additive.

(Note that this does not settle our previous question. Although $\mathcal{E}$ -disintegrability is strictly stronger than $\mathcal{E}$ -conglomerability, to my knowledge, it is an open question whether disintegrability in every countable partition is strictly stronger than conglomerability in every countable partition. So the fact that countable disintegrability entails countable additivity leaves it open whether countable conglomerability entails countable additivity.)

$$\begin{array}{ccc}
\text{Countable Disintegrability} & \implies & \text{Countable Conglomerability} \\
\Downarrow & & \Downarrow \\
\text{CA} & \implies & \text{CA or finitely-valued}
\end{array}$$

Let Finite/Countable Disintegrability be the claim that a credal state C is rational only if $C_{\mathcal{E}}$ is disintegrable in any finite/countable partition $\mathcal{E} \subset \mathcal{F}_C$. The equivalence between Countable Disintegrability and CA suggests the following alternative argument in favor of CA:

P1 Finite Disintegrability is true.

P2 If Finite Disintegrability is true, then Countable Disintegrability is true.

P3 Countable Disintegrability is true if and only if CA is true.

$\therefore$ CA is true.

Call this **the disintegrability argument**.

As before, an immediate worry arises: Any consideration that supports P2 of the disintegrability argument also seems to support the stronger claim that, if Finite Disintegrability is true, then a credal state C is rational only if it is totally disintegrable, i.e. $C_{\mathcal{E}}$ is disintegrable in any partition $\mathcal{E} \subset \mathcal{F}_C$. The question then is whether proponents of the disintegrability argument for CA can accept Total Disintegrability as a rational norm as well, over and above Countable Disintegrability. If they can’t, then the argument would

appear just as dialectically ineffective as its conglomerability counterpart. Relativizing conditional credences to partitions would be an unnecessary detour.

I think the CA theorist can't accept Total Disintegrability as a rational norm. The basic idea is simple: Total Disintegrability rules out intuitively permissible credal states. In particular, let $\Omega = [0, 1]$. It turns out that there are no credal states C such that $C(\cdot|\Omega)$ is the Lebesgue measure and C is totally disintegrable. More specifically, there is a Borel partition $\mathcal{E} \subset \mathcal{F}_C$, whose elements are only singletons or doubletons, such that $C_{\mathcal{E}}$ is disintegrable in $\mathcal{E}$ only if there is some $E \in \mathcal{E}$ such that $C_{\mathcal{E}}(\cdot|E)$ is not even a finitely additive probability measure (Dubins and Prikry 1995, Corollary 1). So if Total Disintegrability is a rational requirement, then it must be rationally impermissible to have a uniform prior over the unit interval. But again, this seems to be a hard pill to swallow, especially for the CA theorist.

6.3 Weak disintegrability for credal states

Now there is a way of weakening the disintegrability requirement to potentially get around this objection. Recall that $C_{\mathcal{E}}$ is disintegrable in $\mathcal{E}$ just in case

$$\forall A \in \mathcal{F}, C(A|\Omega) = \int_{\Omega} C_{\mathcal{E}}(A|\cdot) dC.$$

As mentioned above, this equation holds only if the integral on the right-hand side is well-defined. So disintegrability in $\mathcal{E}$ presupposes that $C_{\mathcal{E}}$ satisfies certain measurability condition, which translates into the requirement that, for any rational r , the agent has well-defined unconditional credence in the proposition that her conditional credences in A given the true partition element would be at least r .

But one might think *this* requirement is unmotivated. Why should it be a requirement of rationality that one has well-formed opinions about how likely one's conditional credence in some proposition given a partition is above certain threshold? This worry is made more salient by the observation that this measurability condition becomes an additional constraint only when the relevant partition is uncountably infinite, i.e. the proposition corresponds to an uncountable disjunction. But insofar as one is allowed to have some credal gaps, why should one be required not to have gaps in those complicated disjunctions specifically?

I think there is something to be said about the measurability condition (cf. Chapter 2). But let me concede the objection here for the sake of argument. Say a conditional credence function $C_{\mathcal{E}}$ is **weakly disintegrable in $\mathcal{E}$** if there is an unconditional credence function $\tilde{C}$ over $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$ such that

1. $\tilde{C}$ extends C , in the sense that for every $\{E_i : i \in I\} \subset \mathcal{E}$, if $\bigcup_i E_i \in \mathcal{F}_C$, then $\tilde{C}\{\{E_i : i \in I\}\} = C(\bigcup_i E_i)$;
2. $\forall A \in \mathcal{F}, C_{\mathcal{E}}(A|\Omega) = \int_{E \in \mathcal{E}} C_{\mathcal{E}}(A|E) d\tilde{C}$.

While $\mathcal{E}$ -disintegrability requires that one's unconditional credence in A be equal to one's unconditional expectation of one's conditional credence in A given the true element of the partition $\mathcal{E}$, weak $\mathcal{E}$ -disintegrability requires that these two be equal, if one *were to* have well-defined unconditional expectation of one's conditional credence in A given the true element of the partition $\mathcal{E}$. These two requirements are equivalent whenever $C_{\mathcal{E}}$ satisfies the measurability condition. So in particular, they are equivalent when the sample space is countable.

This suggests the following amendment of the disintegrability argument:

P1 Weak Finite Disintegrability is true.

P2 If Weak Finite Disintegrability is true, then Weak Countable Disintegrability is true.

P3 Weak Countable Disintegrability is true if and only if CA is true.

$\therefore$ CA is true.

Call this **the weak disintegrability argument**.

Admittedly, the weak disintegrability argument would be no more dialectically effective than its stronger counterpart, if proponents of this argument deny Weak Total Disintegrability, thereby imposing Weak Countable Disintegrability as an arbitrary stopping point. But since Weak Total Disintegrability is strictly weaker than Total Disintegrability, the hope is that proponents of the weak disintegrability argument can accept the former while denying the latter. In particular, one may be hopeful that there is a credal state C such that $C(\cdot|\Omega)$ is the Lebesgue measure over the unit interval and $C_{\mathcal{E}}$ is weakly disintegrable in every Borel partition $\mathcal{E} \subset \mathcal{F}_C$.

Unfortunately, the existence of such a credal state also remains an open mathematical problem (Berti et al. 2017). So this response does not obviously succeed. But I think there is a more pressing worry: As it turns out, Weak Total Disintegrability entails the Continuum Hypothesis, if it is a rational requirement that both unconditional and conditional credences must be countably additive. Even worse, Weak Total Disintegrability also entails the negation of the Continuum Hypothesis, under a different, equally plausible assumption. So I think the CA theorist can't accept Weak Total Disintegrability, which in turn casts doubt on the weak disintegrability argument.

6.4 The Continuum Hypothesis

Let me start with the first horn of the dilemma. Recall that in §1.3, I distinguished between CA and CA-COND. In the Kolmogorovian framework, they correspond to the following two claims:

CA. A credal state C is rational only if $C(\cdot|\Omega)$ is countably additive.

CA-COND. A credal state C is rational only if, for every $\mathcal{E} \subset \mathcal{F}_C$ and $E \in \mathcal{E} \cup \{\Omega\}$, $C_{\mathcal{E}}(\cdot|E)$ is countably additive.

CA-COND follows from CA together with Conditionalization and Closure – the principle that, if one starts out with a rational prior and updates rationally, then one is guaranteed to end up with a rational posterior. I find Closure a compelling principle. To this extent, I think a Bayesian should accept CA-COND if she accepts CA.

However, it turns out that whether Weak Total Disintegrability is compatible with CA-COND depends on the truth of the Continuum Hypothesis. More precisely:

Fact 2 (Dubins and Prikry 1995). *Given ZFC, the following are equivalent:*

1. *There is a credal state C such that*

- *$C(\cdot|\Omega)$ is the Lebesgue measure over the unit interval (with $\mathcal{F}_C$ being the Borel algebra). and*
- *for every Borel partition $\mathcal{E} \subset \mathcal{F}_C$, $C_{\mathcal{E}}$ is weakly disintegrable in $\mathcal{E}$ and $C_{\mathcal{E}}(\cdot|E)$ is countably additive for every $E \in \mathcal{E}$*

2. *the Continuum Hypothesis (CH) is true.*

So if it is rationally permissible to adopt a credal state whose unconditional part is the Lebesgue measure over the unit interval, then any agent with such a credal state can satisfy the joint requirements of Weak Total Disintegrability and CA-COND only if CH is true. In other words, a CA theorist can accept Weak Total Disintegrability as a rational norm only if she (i) accepts CH, or (ii) denies the rational permissibility of having a uniform prior over the unit interval, or (iii) denies that countable additivity is a rational constraint on conditional credences. Since neither (ii) nor (iii) seems a plausible way out, let me focus on (i) in the remainder of this section.

The first thing to note is that this is a surprising development. Pre-theoretically, one might have thought that how we settle the epistemological question whether rational credences should be countably additive, has nothing to do with the ontological question whether there is an uncountable set with cardinality smaller than that of the reals. But proponents of the weak disintegrability argument for CA are pressured to take a stand on the latter question, in order to avoid begging the question against their opponents. Some may view this as a methodological consideration against settling the additivity debate based on disintegrability-related considerations. Epistemology should not be held hostage to mathematical ontology. If the weak disintegrability argument goes through only if CH is true, then so much the worse for the argument.

Now I think the relationship between epistemology and mathematical ontology is much more nuanced. Indeed one might think that *that* is precisely what is at stake. Fundamentally, the debate about countable additivity is a debate about the meaning and significance of infinitary representations in our epistemological theories. How should we understand models that involve credal distributions over infinitely many propositions? Are there distinct infinitary rational norms, over and above the finitary ones? One may not have expected *ex ante* that answering these questions would require one to take a stand on issues like CH. But that is no reason to abandon the investigation. Part of the beauty of philosophical inquiry is that it is full of unexpected twists and turns.

Nevertheless, I think proponents of the weak disintegrability argument should at least balk at accepting CH. This brings us to the second horn of the dilemma: Conglomerability/Disintegrability-based considerations have often been used to argue *against* the truth of CH, rather than in its favor. For instance, Freiling (1986) gave an argument against CH on the basis of the following principle:

If an accurate Yes-No prediction can always be made after a preliminary event takes place and, no matter what the outcome of that event, the prediction is always the same, then the prediction is also in some sense accurate before the preliminary event.

This is of course just a special consequence of Total Conglomerability. (Indeed, there is a close parallel between Freiling's argument against CH and the inconglomerability argument against FA – see Appendix A.) So there is at least one tantalizing route from Total Conglomerability (and therefore Weak Total Disintegrability) to the *negation* of CH.

Moreover, while many philosophers think credal gaps are rationally permissible, i.e. rationality does not require one to assign well-defined unconditional credences to every proposition that can be represented by a subset of the relevant sample space, no philosophers to my knowledge think that credal gaps can be rationally required. That is, the following principle seems plausible:

Permissible Totality. For any Ω , it is not a rational requirement that $C(A|\Omega)$ be undefined for any $A \subseteq \Omega$.

However, given Permissible Totality, CH is incompatible with CA:

Fact 3 (Oxtoby 1971, Theorem 5.6). *Fix $\Omega = [0, 1]$. Under ZFC, 1 implies 2:*

1. *there is a credal state C such that*

- $\mathcal{F}_C = \mathcal{P}(\Omega)$;
- $C(\cdot|\Omega)$ is a probability measure and countably additive;
- $C(\{w\}|\Omega) = 0$ for all $w \in \Omega$.

2. the Continuum Hypothesis (CH) is false.¹⁸

Since CA is equivalent to Countable Disintegrability, and Weak Total Disintegrability implies Countable Disintegrability,¹⁹ it follows that Weak Total Disintegrability implies $\neg$ CH, assuming the permissibility of having a prior unconditional credence distribution over $[0, 1]$ that is uniformly zero on the singletons and well-defined on every subset.

$$\begin{array}{ccc} \text{Weak Total Disintegrability} & \xrightarrow{\text{CA-COND}} & \text{CH} \\ \Downarrow & & \\ \text{Countable Conglomerability} & \xrightarrow{\text{Permissible Totality}} & \neg\text{CH} \end{array}$$

To summarize: The Kolmogorovian thinks rational conditional credence should be relativized to background partitions (thought of as questions to which the conditioning propositions are complete answers). While this would block the “question-begging” objection to the original conglomerability argument, it also undermines the logical connection between conglomerability and countable additivity. On the other hand, if we appeal to a (putatively) stronger condition than countable conglomerability – (weak) countable disintegrability, then the resultant argument appears to fall to an analogous version of the question-begging objection: The argument seems dialectically compelling only if the CA theorist also accepts (weak) total disintegrability as a rational norm on credal states. But it appears that she can’t, on pain of denying the rational permissibility of a uniform and/or total prior over the interval.

7 Coda

Let me conclude with two inconclusive remarks. First, in this chapter I have focused primarily on conglomerability/disintegrability principles as *primitive* rational constraints. Pre-theoretically, there seems to be something intrinsically bad about having conditional credences/credal states that are inconglomerable or indisintegrable. The question is whether this pre-theoretical judgment can be made into a precise principle that supports CA.

There is however a different approach. One might think the rationality of conglomerability/disintegrability

¹⁸Hoek (2020a) uses a version of this result (with respect to chance rather than credence) as the basis for an abductive argument against CH. I think Fact 2 puts a wrinkle in his argument. For one might think that (i) if unconditional chances should be countably additive (which is one of the assumptions of Hoek’s argument), then conditional chances should be countably additive as well; (ii) for any partition $\mathcal{E}$, the unconditional chance of a proposition A should equal the unconditional chance expectation of the conditional chance of A given the true element of $\mathcal{E}$. In other words, the analogues of CA-COND and (Weak) Total Disintegrability are true with respect to chance. But then Fact 2 shows that CH must be *true* under these two assumptions (assuming that there is a chance measure over the unit interval whose restriction to the measurable sets equals the Lebesgue measure). In other words, like the CA theorist, proponents of the Hoek-style argument against CH must also concede a discontinuity between Countable Disintegrability and Total Disintegrability: Chances are countably disintegrable (i.e. countably additive), but not totally disintegrable. (For what it’s worth, I find this discontinuity even less acceptable than that with respect to rational credences.)

¹⁹Recall that the measurability condition is automatically satisfied when the partition is countable. So Weak Countable Disintegrability is equivalent to Countable Disintegrability

is grounded in more fundamental norms. Indeed, it would appear that the anti-CA theorist must think so: conglomerability is compelling only insofar as it is a consequence of Coherence. The CA theorist could adopt a similar attitude. There is nothing intrinsically bad about inconglomerability. Rather, conglomerability is a rational requirement only insofar as it is entailed by other norms, such as maximizing expected accuracy or not having one's credences be strictly dominated by an alternative probability measure over a partition (Easwaran 2013a, 2013c). The thought is, if these norms only justify Countable Conglomerability but not Total Conglomerability, then the CA theorist can run the conglomerability argument without begging the question against her opponents.

I am open to this line of investigation. My main reservation is that, we seem to have a stronger intuition about the rational force of conglomerability than these alternative principles. It is not immediately clear that one must always update to maximize expected accuracy. The prospect of having one's credences be strictly dominated by an alternative probability measure also seems less bad, given that it is often permissible if not required to have one's credences be weakly dominated by an alternative probability measure (e.g. the Lebesgue measure is weakly dominated by any Dirac measure). On the other hand, Debby, Lotti and Coorinne all seem irrational at first glance. This is of course not a weighty consideration against the project – the order of intuitiveness does not always track the order of explanations. Nevertheless, I think this consideration does cast doubt on the dialectical effectiveness of arguments based on these alternative principles.

Second, Seidenfeld et al. (2014) conclude from their result that

[T]he anomalous phenomenon of nonconglomerability is a result of adopting the de Finetti/Dubins theory of coherent conditional probability instead of the rival Kolmogorovian theory of regular conditional distributions. It is not a result of the associated debate over whether probability is allowed to be merely finitely additive rather than satisfying countable additivity."

I think my discussion both supports and qualifies their insight. On the one hand, both the CA theorist and her opponent have to allow some form of inconglomerability, *whether or not* they are coherentist or Kolmogorovian about rational conditional credences. So the question that is relevant to the additivity debate is not *whether*, but *when*, can rational agents be inconglomerable.

On the other hand, I think our discussion in §6 suggests that the question of conglomerability is not entirely settled by one's preferred theory of conditional probability either. For although Kolmogorov's theory bakes conglomerability into the definition of rational credal state (via the integral formula), there are cases where no such credal states may exist under the constraint of countable additivity (the case for finite additivity remains unclear). To this extent, the relationship amongst the three issues – conglomerability,

conditionalization and countable additivity – is more nuanced than suggested by the quote above. Perhaps the general philosophical takeaway is what I noted at the beginning: One must be careful when treading in the land of infinities. For infinities are weird, and some are weirder than others.

Chapter 2

Closure and Countable Additivity

1 Introduction

The previous chapter argued that the conglomerability argument for CA does not succeed. For proponents of that argument must draw a normative distinction between inconglomerability in countable and uncountable partitions, and yet it is unclear what could justify that distinction, other than the presupposition that CA is a rational requirement.

Indeed, the pattern seems to generalize. For instance, consider the Dutch book argument. If an agent has credences that are finitely but not countably additive, then, assuming that credences equal to betting quotients, this agent would be disposed to accept a countable infinity of bets that jointly guarantee them a sure loss (Williamson 1999). In response, the anti-CA theorist would point out that the same is true for any cardinality. Given an infinite cardinal κ , an agent whose credences are not κ -additive would be disposed to accept a set of bets of cardinality κ that jointly guarantee them a sure loss. Proponents of the Dutch book argument for CA must therefore draw a normative distinction between susceptibility to countably infinite Dutch books and susceptibility to uncountably infinite Dutch-books – one reveals irrationality, the other one does not. But again, it is unclear what could justify that distinction, other than the presupposition that CA is a rational requirement.¹

At this point, the debate seems to have hit a stalemate. On the one hand, the CA theorist could dig in their heels and insist that something is better than nothing. Even if you can't save everyone, it doesn't follow

¹As Easwaran (2013c) points out, if we impose a stronger condition on Dutch books – a Dutch book must consist of bets that are not only acceptable but *favorable*, in the sense that they have strictly positive expected payoffs – then only failures of countable additivity would be Dutch-bookable. One wrinkle with his argument is that it exploits the structure of the reals; in particular, the sum of uncountably many positive reals is infinite. But one might think, as we move to higher cardinalities, we should allow the mathematical structure of the agent's preference to be similarly complex (e.g. her utilities may be representable by hyperreals rather than reals). Whether this would block Easwaran's argument is a question for future research.

that you should save no one. On the other hand, the anti-CA theorist would remind the CA theorist that countable additivity does not come without costs. If CA is a rational norm, then many intuitively plausible credence functions would be ruled out as irrational, including uniform distributions over the natural numbers, or shift-invariant distributions over the powerset of the unit interval (Finetti 1970; Savage 1954).

This paper seeks to break the stalemate. My strategy is to look at the additivity debate from a diachronic perspective. In particular, I propose the following bridge principle between synchronic and diachronic rationality:

Closure. Rational updating preserves rationality. More specifically, if you start out with a rational prior and update in accord with a rational updating plan, then you are guaranteed to end up with a rational posterior.

I show that, assuming Closure, CA is inconsistent with very weak rational constraints. To this extent, it is an even stronger axiom than has been appreciated in the literature.

Here is the plan: §2 introduces the principle of Closure. §3 gives the first Closure-based argument against CA. §4-5 give the second argument and consider some possible replies. §6 concludes with more general philosophical reflections.

2 Closure

At first glance, Closure seems almost like a tautology: A rational updating plan is, *by definition*, a procedure that preserves rationality.² Indeed, while to my knowledge the principle has not been stated explicitly in the literature, it is often implicitly assumed in the background. For instance, Lewis (1980) appeals to a version of this principle when giving a definition of reasonable priors:

[A credence function] is to be reasonable in the sense that if you started out with it as your initial credence function, and if you always learned from experience by conditionalizing on your total evidence, then no matter what course of experience you might undergo your beliefs would be reasonable. (p.271)

To give another example: Some philosophers think that rational credence should be *regular*, i.e. rational agents should not assign zero credence to contingent logical/metaphysical possibilities. One objection to

²In fact one might think a stronger principle is true: you should always plan to adopt a rational posterior, irrespective of the rationality of your prior. Everything I say in this paper applies to this stronger principle *a fortiori*. For what it's worth, I myself am skeptical of the stronger principle. Suppose you are fully confident that it will rain and it will not rain tomorrow, and you learn that a fair coin landed heads. Should you revise your credence function to be probabilistically coherent? If so, how confident should you be that there will be rain tomorrow? I think the answer is not obvious, and our intuitions are pulled in different directions by different considerations (e.g. the principle of total evidence, the principle of conservatism, the norm of accuracy, etc.). See Staffel (2019, ch.7) for a discussion about the complications involved in evaluating credal change for agents with irrational priors.

Regularity as a rational norm is that the set of regular probability functions is not closed under Bayesian conditionalization, i.e. C is regular $\nRightarrow C(.|E)$ is regular. In fact, $C(.|E)$ would be irregular whenever E is contingent.³

This objection again presupposes Closure. The argument is: Assuming the rationality of Bayesian conditionalization, $C(.|E)$ must be rational if C is rational. But Regularity implies that $C(.|E)$ is not rational even if C is rational. So the principle must be false as a universal rational norm.⁴

Despite being close to a truism, Closure is still subject to potential counter-examples. One class of such cases involves updating on misleading higher-order evidence – evidence that suggests (falsely) that your response to your total evidence is irrational. For instance, suppose you have good grasp of basic arithmetical and calendar facts. You are also correctly certain of the following interesting trivia: the date 123456789 days from June 1, 2021 is Aug 16, 340034. Let this proposition be p . Now suppose a mathematician friend of yours, whom you respect deeply, tells you that she thinks differently: the date should be Aug 15, 340034. Some philosophers think in this case, you should become less than fully confident in p , even though doing so would make your credences incoherent (since p is logically entailed by other things that you are certain of). This is the peculiar feature of misleading higher-order evidence: responding to such evidence rationally leads you from rationality to (putative) irrationality.⁵

The question how to respond rationally to higher-order evidence is complicated and lies beyond the scope of this paper (it is also unclear whether the above example should be treated as a counter-example to Closure or Probabilistic Coherence). But fortunately, we can set aside this question for now. For the cases I am going to discuss will all be “vanilla” cases that only involve the learning of some piece of true, first-order empirical information. Even if Closure is false, it appears that a restricted version of the principle should hold for cases of this kind.

So far I have tried to motivate Closure on an intuitive level. But in order to use that principle to do any work, we need to make it more precise. In particular, the principle invokes the notion of an “updating plan.” But what is an updating plan? Let us turn now to this question.

2.1 Updating plans

Let’s focus on plans for the moment. We plan on a daily basis. I make plans about what to have for breakfast, what to read for the summer, etc. Many of our plans take the form of a family of conditionals

³See Huttegger (2017, ch 5.2) for a discussion of this objection. (Huttegger attributes this objection to Richard Jeffrey. Earman (1992, p.42) discusses a similar objection.

⁴It is worth noting that Lewis, who accepts both regularity and conditionalization as the only rational updating plan, escapes the trilemma by restricting the scope of regularity to initial credence functions only.

⁵See Worsnip (2018) for a detailed discussion on the idea that misleading higher-order evidence creates a rational dilemma where responding to your evidence leads to violations of inter-level coherence.

whose antecedents are mutually exclusive and jointly exhaustive (i.e. they form a partition of the space of possibilities). For instance, I may plan to have cereal for breakfast if I have milk left in my fridge, and toast otherwise. Abstractly, we can represent a plan of this kind by a function

$$f : \mathcal{E} \rightarrow \mathcal{A}$$

where $\mathcal{E}$ is a partition and $\mathcal{A}$ is a set of actions. In my breakfast example, my plan corresponds to the function $f : \{\text{milk, no milk}\} \rightarrow \{\text{have cereal, have toast}\}$ with

$$f(\text{milk}) = \text{have cereal}, \quad f(\text{no milk}) = \text{have toast}.$$

Conceptually, updating plans are plans of a special sort. Instead of telling you what to do, they specify what credence function you should adopt, given each piece of evidence that you might obtain. Formally, an updating plan is represented by a function f that maps evidence to credence functions – the functions that you plan to adopt, if you learn the respective pieces of evidence.

Let me pause here to make a technical remark about credence functions. In the previous chapter, I treated conditional credences as primitive; a credence function is a *binary* function that assigns real numbers to pairs of propositions. For the purposes of this chapter, I shall treat unconditional credences as primitive, and represent a credence function by a *unary* function that assigns real numbers to individual propositions. The choice is merely for notational convenience. In particular, as I discuss in §2.2, assuming Conditionalization, there is a 1-1 correspondence between conditional credence functions and pairs of unconditional credence functions and updating plans.

Let Ω be a sample space and C an unconditional credence function over Ω . (As before, propositions are represented as subsets of Ω .) Let $\Delta(\Omega)$ denote the set of credence functions over Ω . Each $Q \in \Delta(\Omega)$ is interpreted as the cognitive act of adopting credence function Q . Finally, let $\mathcal{E}$ be a collection of evidence. For present purposes, we shall assume that evidence is propositional, factive and transparent, i.e. $\mathcal{E}$ is a partition of $\mathcal{P}(\Omega)$.

An **updating plan given $\mathcal{E}$ and C** is represented by a function

$$f_C : \mathcal{E} \rightarrow \Delta(\Omega)$$

with the intended interpretation: If E is your total evidence, then revise your credence function from C to $f_C(E)$.⁶

⁶For discussions of updating plans, see Greaves and Wallace (2005), Schafer (2014), Easwaran (2013a), Schoenfield (2018),

It may be helpful to work through a concrete example. Suppose you are about to learn whether it will rain tomorrow. You know your source is reliable. Let C be your initial unconditional credence function (defined over some possibility space Ω), and R be the proposition that there will be rain. Suppose your initial credences for R and $\neg R$ are well-defined and positive. Here is one possible updating plan: if you learn that there will be rain, update your credence by conditionalizing on R ; if you learn that there will not be rain, update your credence by conditionalizing on $\neg R$. Formally, this plan can be represented by the function $f_C : \{R, \neg R\} \rightarrow \Delta(\Omega)$ with

$$f_C(R) = C(.|R), \quad f_C(\neg R) = C(.|\neg R)$$

where $C(.|E) = \frac{C(. \& E)}{C(E)}$.

We are now ready to state Closure more precisely:

Closure. For any rational updating plan f_C given $\mathcal{E}$, if C is rational, then $f_C(E)$ is rational for every $E \in \mathcal{E}$.

Before moving on, let me give a schematic sketch of the arguments to come.

2.2 Closure and CA

The main focus of my arguments will be the following consequence of CA and Closure:

Additivity Closure. Rational updating plans preserve countable additivity. More precisely: for any rational updating plan f_C given $\mathcal{E}$, if C is rational, then for every $E \in \mathcal{E}$, $f_C(E)$ is also countably additive.

Both of my arguments will be arguments by counter-examples. More specifically, they take the following schematic form:

P1 In situation S, it is rationally permissible to follow an updating plan f that has properties $\Phi_1, \dots, \Phi_n$.

P2 In situation S, any updating plan f that has properties $\Phi_1, \dots, \Phi_n$ fails to preserve countable additivity, i.e. there is E such that $f(E)$ is not countably additive.

$\therefore$ Additivity Closure is false in S. Assuming Closure, it follows that CA is not a universal rational norm.

Two remarks before moving on. First, strictly speaking, the entailment holds only under the assumption that, in situation S, there is at least one rational updating plan that the agent could adopt. One could

and Gallow (2019). In particular, following Easwaran (2013a) and Gallow (2019), I assume that agents do not need to have voluntary control over their doxastic states in order to carry out updating plans. Instead, it suffices that they manifest relevant updating *disposition* in response to their evidence (cf. Chapter 4 and 5).

argue that, in the examples that I am about to discuss, the agent's epistemic situation is so complex that it lies beyond the reach of diachronic rationality. While I don't think this is a plausible response to my first argument, it may have some attraction as a response to my second argument. I discuss it in §6.

Second, both of my arguments target Additivity Closure as a constraint on rational updating plans. But one could also read them as arguments against CA-COND, interpreted as a constraint on rational conditional credences:

CA-COND. Conditional credences are countably additive. More precisely: For any rational conditional credence function C , $C(.|E)$ is countably additive whenever it is well-defined.

CA-COND follows from Additivity Closure and Conditionalization (or equivalently CA Closure and Conditionalization). Moreover, given Conditionalization, Additivity Closure and CA-COND come apart only when the conditioning proposition receives zero unconditional credence. The reason why I prefer to focus on Additivity Closure instead of CA-COND is two-fold. First, as we saw from the previous chapter, it is controversial what conditionalization (or rational conditional credence) even means when the conditioning proposition is null.⁷ More importantly, my second argument relies on principles that are much weaker than Conditionalization. So the argument has a bite, even on non-Bayesians who think rational agents may update in accord with rules other than conditionalization.

3 First argument

Let's start with a simple example. Suppose someone draws a ball randomly from an urn. The only thing you know with certainty is that the urn contains balls of two colors exclusively: red and green. How confident should you be that the ball drawn from the urn is red, rather than green?

Intuitively, the answer seems to be 0.5. After all, there are only two possibilities, and you have no reason to think that one is more likely than the other. (It may also be worth noting that the problem of partition-dependence, which many take to be the strongest objection to the Principle of Indifference, does not arise in this case. For there is only one non-trivial partition of two atomic possibilities.)

But for the sake of argument, let's grant that you are not required to split your credence in this case. In fact, let's grant that you are not required to have any particular credence. Rather, the only assumption I'll

⁷In particular, some (e.g. Kolmogorov) think that in those circumstances rational conditional credence should be undefined without the specification of a partition, i.e. for them, the proper CA-COND principle should be:

CA-COND*. Conditional credences are countably additive. More precisely: For any rational credal state C and partition $\mathcal{E} \subset \mathcal{F}_C$, $C_{\mathcal{E}}(.|E)$ is countably additive whenever it is well-defined.

Like CA-COND, CA-COND* is also a consequence of Additivity Closure (assuming Conditionalization). As we shall see, my first argument works against CA-COND, whereas my second argument works against CA-COND*.

need is that, given your epistemic situation, it is *permissible* for you to be uncertain whether the ball is red or green.

More generally, consider the following principle (cf. Builes, forthcoming):

Weak Regularity. If you learn with certainty that one of two disjoint possibilities is true, and you have no reason to think that one is more likely than the other, then you are not required to rule out one in favor of the other.

Weak Regularity seems quite plausible. To deny Weak Regularity is to say that rationality sometimes mandates *a priori* bias in the absence of reasons. But that seems absurd. How could it be a rational requirement to be 100% sure that one possibility is true, when one's evidence is symmetric between the two?

However, it turns out that, assuming the coherent theory of conditional probability, Weak Regularity is incompatible with CA, given Closure and Conditionalization (equivalently, it is incompatible with CA-COND). The conflict can be illustrated through the following example:

Random Point. Levita picks a nonnegative real number smaller than 1. You are asked to guess what it is. Fortunately, Levita will give you a bit of help by telling you a subset that contains the number she picked. *A priori* you don't know what she is going to tell you. She could be mischievous and tell you that the number is drawn from $[0, 1)$, which you already know. Or she might simply reveal to you the number she picked.

Let $\Omega = [0, 1)$, where each $x \in \Omega$ represents the possibility that the number picked by Levita is x . Suppose your prior credal state is representable by a coherent probability function⁸

$$C(\cdot|\cdot) : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \setminus \{\emptyset\} \rightarrow [0, 1)$$

with $\mathcal{F}_C \subset \mathcal{P}(\Omega)$ being the sigma-algebra generated by rational intervals (also known as the Borel sigma-algebra).

Fact 1. *C is weakly regular only if $C(\cdot|E)$ is not countably additive for some $E \in \mathcal{F}$.*

Proof. Suppose C is weakly regular. We show that there is a countable set $E \in \mathcal{F}$ such that $C(x|E) = 0$ for

⁸Recall that a coherent probability function is a binary partial function $C(\cdot|\cdot) : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \rightarrow [0, 1]$ such that

- $\forall B \in \mathcal{F}_C \setminus \{\emptyset\}$, $C(\cdot|B)$ is a well-defined finitely additive probability function with $C(B|B) = 1$ (also recall that, given a binary credence function C , $\mathcal{F}_C \subset \mathcal{P}(\Omega)$ denotes the set of propositions that receive well-defined unconditional credences under C);
- (multiplicative axiom) $\forall F, G, H \in \mathcal{F}_C$ with $H \neq \emptyset$ and $G \cap H \neq \emptyset$,

$$C(F \cap G|H) = C(F|G \cap H)C(G|H).$$

all x . We construct this set as follows: Pick $x_0 \in [0, 1)$. Let

$$S_n^0 = \{y \in [0, 1) : \frac{C(x_0|x_0 \vee y)}{C(y|x_0 \vee y)} \leq n\}.$$

Note that $[0, 1) = \bigcup_{n \geq 1} S_n^0$. Since $[0, 1)$ is uncountably infinite, it follows that there is n_0 such that $S_{n_0}^0$ must be uncountably infinite as well. Let A_0 be this set subtracting x_0 . Then for every $y \in A_0$, $C(x_0|x_0 \vee y) \leq n_0/(n_0 + 1)$.

Now fix $x_1 \in A_0$. Let

$$S_n^1 = \{y \in A_0 : \frac{C(x_1|x_1 \vee y)}{C(y|x_1 \vee y)} \leq n\}.$$

Again, $A_0 = \bigcup_n S_n^1$. Since A_0 is uncountably infinite, there is n_1 such that $S_{n_1}^1$ is also uncountably infinite. Let A_1 be this set subtracting x_1 . For every $y \in A_1$, we have $C(x_0|x_0 \vee y) \leq n_0/(n_0 + 1)$ and $C(x_1|x_1 \vee y) \leq n_1/(n_1 + 1)$.

Repeat this procedure. Let $E = \{x_0, x_1, \dots\}$ and $E_i = \{x_j : j \geq i\} \subset E$. By construction, for each x_i , and every $j > i$, we have

$$C(x_i|x_i \vee x_j) \leq \frac{n_i}{n_i + 1}.$$

Note that this means that, for any k ,

$$C(x_i|x_i \vee x_{i+1} \vee \dots \vee x_{i+k}) \leq \frac{n_i}{n_i + k}.$$

So for any ϵ , there is k large enough such that $C(x_i|x_i \vee x_{i+1} \vee \dots \vee x_{i+k}) < \epsilon$. It follows that for all $\epsilon > 0$,

$$C(x_i|E) \leq C(x_i|E_i) < \epsilon.$$

So $C(x_i|E) = 0$. Since E is countable, $C(\cdot|E)$ is not countably additive. □

To complete the reasoning: By Weak Regularity (and the existence assumption), there is at least one rationally permissible C such that C is weakly regular. On the other hand, by Additivity Closure, C is rational only if $C(\cdot|E)$ is countably additive for every $E \in \mathcal{F}$, which, by Fact 1, means that C is rational only if C is not weakly regular. A contradiction.

Two quick remarks. First, the proof of Fact 1 does not rely on any assumption about the shape of $C(\cdot|\Omega)$ (though it is not hard to check that Weak Regularity entails that $C(\cdot|\Omega)$ must be uniformly zero on the singletons). In particular, $C(\cdot|\Omega)$ could be any continuous probability distribution. To this extent, the result is very general: Almost all probability distributions used in science and statistics must violate either Weak

Regularity, or fail to be countably additive when conditionalized on some countable subset of their domain.

Second, the proof of Fact 1 actually supports a stronger conclusion: For any coherent C over an uncountable Ω , if C is weakly regular over some uncountable subset S of Ω such that $C(S) > 0$, then $C(\cdot|E)$ is not countably additive for some $E \in \mathcal{F}_C$. Thus, the CA theorist who accepts Additivity Closure and the coherent theory of conditional probability must think that Weak Regularity is false in all but exceptional circumstances. At this point, the pill seems harder and harder to swallow.⁹

Now I think Fact 1 is bad news for CA, but it is not quite decisive, for two reasons. First, the CA theorist must have already made peace with the idea that it can be a rational requirement to have a prior C such that

$$C(w|w \vee v) > 1 - \epsilon, \text{ for some arbitrarily small } \epsilon$$

even in the absence of evidence in favor of w vis-a-vis v . More concretely, let Ω be a countably infinite sample space and C be a prior (e.g. we may suppose that Levita only picks rational numbers between 0 and 1). Then $C(\cdot|\Omega)$ is countably additive only if $C(\{w\}|\Omega) > 0$ for some $w \in \Omega$. Suppose $C(\{w\}|\Omega) = r$. Another consequence of $C(\cdot|\Omega)$ being countably additive is that, for any arbitrarily small $\epsilon > 0$, there is a finite set $S \subset \Omega$ such that $C(S|\Omega) > 1 - \epsilon$.

Now let $\epsilon > 0$. Choose $\delta > 0$ such that

$$\epsilon > \frac{\delta}{\delta + 1}.$$

Let S be a finite set such that $w \in S$ and $C(S|\Omega) > 1 - r\delta$. It follows that, for any $v \notin S$ (and there are infinitely many of them), $C(\{v\}|\Omega) < r\delta$, which means that

$$C(w|w \vee v) \geq \frac{r}{r + r\delta} = \frac{1}{1 + \delta} = 1 - \frac{\delta}{1 + \delta} > 1 - \epsilon.$$

Moreover, the reasoning generalizes to any countably infinite subset of Ω that has positive unconditional probability. So even without Weak Regularity, CA alone entails that, for any countable $\Omega' \subset \Omega$ with $C(\Omega') > 0$ and any arbitrarily small $\epsilon > 0$, there are countably infinitely many pairs $\{w, v\} \subset \Omega'$ such that rationality requires that $C(w|w \vee v) > 1 - \epsilon$. Of course, being required to have $C(w|w \vee v) > 1 - \epsilon$ is not quite the same as being required to have $C(w|w \vee v) = 1$. But the two are not a million miles away either.

There is another way in which the CA theorist could get around Fact 1. The proof of Fact 1 relies heavily on the assumption that C is a coherent conditional credence function. In particular, it satisfies the following principle:

⁹The reader might wonder how Fact 1 relates to another inconsistent triad involving Weak Regularity proved in Builes (forthcoming). I address this question in Appendix B.

Multiplicative Axiom. $\forall F, G, H \in \mathcal{F}_C$ with $G \cap H \neq \emptyset$, $C(F \cap G|H) = C(F|G \cap H)C(G|H)$.

However, as mentioned in Chapter 1, some theorists reject this assumption. Rational conditional credence, they argue, is relativized to the choice of a partition. Sometimes, one may be required to adopt different conditional credence functions given the same proposition, viewed as a member of different partitions.

I discussed one motivating example for this proposal in Chapter 1. Here is another example due to Seidenfeld et al. (2014) (adapted to the present context):

Random Point, cont'd. Same set-up as before, except that this time, you do know beforehand what Levita will tell you: She will tell you the value of Y , which is computed from the number X that she picked based on the following function:

$$Y = \begin{cases} X + X/9, & \text{if } X < 0.9 \\ 10X - 9, & \text{if } 0.9 \leq X < 1. \end{cases}$$

Given the value of Y , how confident should you be that X is smaller than 0.9?

Without loss of generality, let's suppose your unconditional credence over Ω is given by the Lebesgue measure. Here is a piece of intuitive reasoning. Fix y . Suppose Levita tells you instead that Y is within ϵ -neighbourhood of y . Then for any $\epsilon > 0$,

$$C(X < 0.9|Y = y \pm \epsilon) = 0.9.$$

But $C(X < 0.9|Y = y)$ is just a limiting case of $C(X < 0.9|Y = y \pm \epsilon)$. So it appears that, given any value of Y , you should be highly confident that X is smaller than 0.9.

The catch is that the same reasoning leads to a different conclusion if Levita uses a different function. For instance, let

$$Z = \begin{cases} 0 & \text{if } X = 0 \\ 1 - 2X & \text{if } 0 < X < 0.5 \\ 2X - 1 & \text{if } 0.5 \leq X < 1. \end{cases}$$

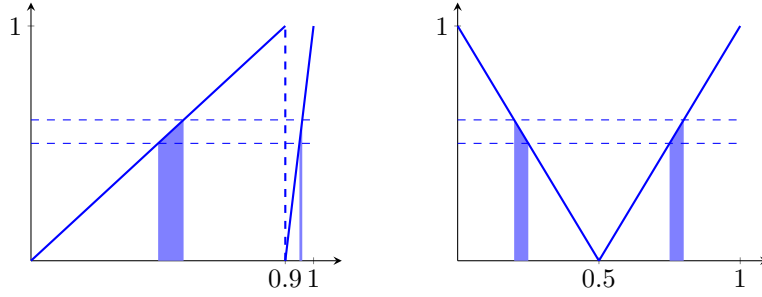
One can check that, for any z and $\epsilon > 0$, $C(X < 0.5|Z = z \pm \epsilon) = 0.5$. So if the previous argument is any good, then you should be 0.5 confident that $X < 0.5$, given any value of Z .

Now consider $Y = 0.1$. This is true if and only if $X = 0.91$ or $X = 0.09$, which in turn correspond to the

case where $Z = 0.82$. So

$$C(X < 0.9|Y = 0.1) = C(X < 0.9|Z = 0.82) = C(X = 0.09|X = 0.91 \vee X = 0.09) = C(X < 0.5|Z = 0.82).$$

But our first argument says that $C(X < 0.9|Y = 0.1) = 0.9$, whereas our second argument says that $C(X < 0.5|Z = 0.82) = 0.5$. A contradiction.



As promised before, I shall discuss the issue of partition-relativity in Chapter 3. At this point, let's grant both complaints. In the next section, I develop a different argument against CA that presupposes neither Weak Regularity (or variants thereof) or the Multiplicative Axiom.

4 Second argument

Let's return to the urn example. Again, suppose someone draws a ball randomly from an urn. You know with certainty that the ball is either red or green. Suppose you also know (with certainty) that the person will tell you the true color of the ball. Question: How confident should you be that the ball is red, if the person tells you that it is red?

The answer seems obvious: If the person tells you that the ball is red, and you know that they are telling the truth, then you should be fully confident that the ball is red. Everything equal, rationality requires that you believe the truth, when you know what it is.

Let's make this intuition a little more precise. Fix a sample space Ω . Suppose you will undergo a learning experience as a result of which, at each state $w \in \Omega$, you will learn a piece of propositional information $E(w)$ and nothing else. For instance, you might look outside of your window, and as a result learn whether or not it is raining. We say your learning experience is **factive** if what you learn will always be true of your actual state, i.e. for all $w \in \Omega$, $w \in E(w)$.

Our previous reasoning motivates the following principle:

Evidential Certainty. If you know that you will undergo a factive learning experience, then you should plan to update by assigning full credence to what you learn, i.e. an updating plan f is rational for you

only if it satisfies the following constraint:

$$\forall E \in \mathcal{E}, f(E)(E) = 1, \text{ where } \mathcal{E} = \{E(w)\}_{w \in \Omega}.$$

(Recall that $f(E)$ is the credence function recommended by f if you learn E and nothing else.)

As noted above, Evidential Certainty seems intuitively plausible. It is also supported by pragmatic and alethic considerations. For instance, if you think E might be true (i.e. $C(E) > 0$) and yet remain uncertain about E after learning that it is true, then you can be Dutch-booked and accuracy-dominated; you are also performing suboptimally in terms of maximizing expected accuracy and pragmatic utility (cf. Chapter 5).

I'll need two principles for my argument. Evidential Certainty is one of them; here is another. As before, suppose you will learn the true color of the ball. Moreover, you also know that (i) you will update on what you learn by Bayesian conditionalization, and (ii) your learning experience will be genuine and transparent, in the sense that you won't lose any of your present knowledge and, for any p , if the person tells you p , then you will know that they've told you p . Formally, the second assumption together with factivity entails that your learning experience is representable by a partitional mapping, i.e. for every w, v , either $E(w) = E(v)$ or they are mutually incompatible. Question: How confident do you expect yourself to be that the ball is red, after you update on the person's testimony?

This time the answer may not seem so obvious, but here is a piece of intuitive reasoning. Suppose you know that you will be fully confident that the ball is red, after you update on the person's testimony. By stipulation, you know that this could only happen if the person will tell you (truthfully) that the ball is red. But conditional on that piece of information, you should be confident *now* that the ball is red. Compare: If you are struggling with a multiple-choice question and you know that you will choose C once you see the answer sheet, then there is no point in waiting – you should choose C now, unless you are a fan of making mistakes.

The previous argument justifies what is known as the Reflection principle: Your conditional credence in p , given that you will be x confident in p , should equal x . While the Reflection principle is found dubious by many due to counter-examples involving memory loss or anticipated future irrationality, most theorists agree that it is plausible under the kinds of assumptions mentioned above. Indeed, Briggs (2009) prove that, given those assumptions, one can derive the Reflection principle as a consequence of Bayesianism.

For the purposes of my argument, however, I will only need the following consequence of Reflection:

Weak Reflection. If you know that you will undergo a learning experience that is factive, genuine and transparent, and you will update on what you learn in accord with a rational updating plan, then for any

proposition A in the domain of your prior, your prior confidence in A should equal your prior expectation of your posterior confidence in A . More precisely: Given a rational C and a learning experience representable by $\mathbf{E}$, an updating plan f given C and $\mathcal{E}$ is rational only if

$$\forall A \in \mathcal{F}_C, C(A) = \mathbb{E}_C[\tilde{f}(A)]$$

where $\tilde{f}$ is the random variable that maps each world $w \in \Omega$ to the unique credence function $f(\mathbf{E}(w))$.

Since violations of Reflection are Dutch-bookable (including in the infinite context – see Rescorla (2018)), it follows that violations of Weak Reflection are Dutch-bookable as well.

Let's take stock: I've introduced two principles for rational updating plan: Evidential Certainty and Weak Reflection. Taken together, they say that, in the most vanilla learning situation (i.e. it is factive, genuine and transparent), rational agents who know that they are rational should (i) plan to be certain of what they learn, and (ii) have their prior align with their prior expectation of their (planned) posterior.

Unfortunately, it turns out that this conjunction can't always be true, if rational updating plans are required to preserve countable additivity.

Rational Difference. Same set-up as before: Levita picked a number from the unit interval at random; you are asked to guess what it is. This time, she will help you in the following way: She will tell you the unique set that contains the number she picked, as well as all the other numbers (between 0 and 1) that differ from that number by a rational. For instance, if she picked 0 (or any other rational number for that matter), then she will tell you that she picked a rational. You know this beforehand.

Let $\Omega = [0, 1]$ and $\mathbf{E}(x) = \{y \in \Omega : |y - x| \in \mathbb{Q}\}$. Let $\mathcal{E} = \{\mathbf{E}(x) : x \in \Omega\}$, which represents the set of all possible information that you will learn from Levita.

Fact 2 (Dubins 1977). *If C is the Lebesgue measure over $[0, 1]$, then an updating plan f given C and $\mathcal{E}$ satisfies Evidential Certainty and Weak Reflection only if, for almost every $E \in \mathcal{E}$, $f(E)$ is not countably additive.*

Thus, if Evidential Certainty and Weak Reflection are genuine rational norms, then in cases like Rational Difference, a rational agent is *required* to violate CA in almost all but exceptional circumstances. So CA is incompatible with Evidential Certainty and Weak Reflection, given Closure. (In fact, it is incompatible with the rational permissibility of satisfying both requirements at the same time, given Closure.) Moreover, this is so, even if we relativize your updating plan to a partition. You knew in advance that you will learn a member of $\mathcal{E}$, and there is no updating plan that satisfies the conditions of Evidential Certainty and Weak Reflection while preserving countable additivity.

There are three possible ways in which a CA theorist could reply to Fact 2: (i) deny Evidential Certainty; (ii) deny Weak Reflection; (iii) deny the coherence or possibility of the case. Let me discuss each of them in turn.

5 Replies

5.1 Evidential Certainty

Recall that Evidential Certainty says: If you know that what you learn will be true, then you should plan to be fully confident in what you learn. As noted above, violation of Evidential Certainty seems both intuitively irrational and instrumentally suboptimal.

However, there is a caveat: All the arguments for Evidential Certainty mentioned above go through, only if the agent assigns positive prior credence to every proposition that she might learn as a result of her learning experience. Take the Dutch book argument for example. Intuitively, violation of Evidential Certainty is Dutch-bookable (assuming factivity) because, if you learn that E but assigns credence $p < 1$ to E , then a bookie who will undergo the same learning experience could exploit you as follows: If both of you learn E , buy a bet from you that costs $\$p$ and pays $\$1$ if E is true; otherwise, do nothing. The bookie can then secure a sure profit by first selling you a side bet against the event of learning E , i.e. you would pay him some money, and get a reward when you actually learn E (and the reward is small enough such that it does not offset your loss from the later bet). However, if you are confident that you will *not* learn E , then no such side bet would be acceptable by your lights. As a result, no bookie can make you suffer a sure loss in this case, *despite* your violation of Evidential Certainty.¹⁰

On the other hand, in Rational Difference, if you start out with a uniform prior, then for every set S that Levita might reveal to you as the set that contains the number she picked, before she passes on that information, you would be certain *ex ante* that she did not pick a number from S . This is because every such set S is countably infinite (for there are only countably many real numbers that differ from a particular real by a rational), and the Lebesgue measure of any countable set is zero.

So one might think Evidential Certainty is too strong after all. The principle seems plausible when the agent assigns positive credence to all her possible evidence. But, in cases like Rational Difference where the agent is planning for (uncountably) infinitely many contingencies, each of which seems improbable *a priori*, we should perhaps cut her some slack.

¹⁰However, one *is* Dutch-bookable if one violates Evidential Certainty with respect to a set of evidence whose disjunction receives positive probability, i.e. if one violates Subjective Evidential Certainty (see below).

I am sympathetic to this complaint.¹¹ The question, however, is how much slack we should cut. At first glance, the following seems to be a plausible alternative to Evidential Certainty:

Subjective Evidential Certainty. If you know that you will undergo a factive learning experience, then it is a rational requirement that you be confident that you will assign full credence to what you learn. Formally: $C(\bigcup\{E \in \mathcal{E} : f(E)(E) = 1\}) = 1$ ¹²

While Evidential Certainty requires you to be certain of what you learn (assuming that your evidence is factive), Subjective Evidential Certainty only requires that you be certain that you will be certain of what you learn. In particular, in Rational Difference, if your prior is the Lebesgue measure, then Subjective Evidential Certainty says that, for countably infinitely many possible S , it is permissible for you to not be fully confident that Levita picked a number from S , even if that is what she tells you explicitly. This seems to be a reasonable concession. Moreover, while there are no Dutch-book arguments for Evidential Certainty, there is a Dutch-book argument for Subjective Evidential Certainty (Meehan and Zhang (forthcoming)).

So the hope is that one may get around Fact 2 by weakening Evidential Certainty to Subjective Evidential Certainty. Unfortunately, this will not work. In fact, it turns out that CA is not even compatible with the combination of Weak Reflection and the following, even weaker principle:

Subjective Evidential Trust. If you know that you will undergo a factive learning experience, then it is a rational requirement that you are *not* confident that you *won't* assign full credence to what you learn. Formally:

$$C(\bigcup\{E \in \mathcal{E} : f(E)(E) = 0\}) < 1$$

Fact 3 (Seidenfeld et al. 2001). *If C is the Lebesgue measure over $[0, 1]$, then*

1. *an updating plan f given C and $\mathcal{E}$ satisfies Subjective Evidential Certainty and Weak Reflection only if $f(E)$ is not countably additive for almost every $E \in \mathcal{E}$;*
2. *an updating plan f given C and $\mathcal{E}$ satisfies Subjective Evidential Certainty and Weak Reflection only if $f(E)$ is not countably additive for some $E \in \mathcal{E}$.*

But Subjective Evidential Trust is extremely weak. It only requires that you be less than fully confident that you will assign full credence to the *negation* of your evidence. Even those who think we never learn anything

¹¹Though it's worth bearing in mind that Fact 2 shows that CA is incompatible with the joint *permissibility* of satisfying Evidential Certainty and Weak Reflection. So, if the CA theorist accepts Weak Reflection as a rational requirement, then she must deny that, in Rational Difference, it is even rationally permissible to be certain of what you will learn. That sounds to me much less plausible. It is one thing to say that one can occasionally break the rule, quite another to say that one must sometimes break the rule.

¹²Note that, in order to satisfy Subjective Evidential Certainty, the set $\{w : \forall E \in \mathcal{E}, w \in E \rightarrow f(E)(E) = 1\}$ must be a member of $\mathcal{F}_C$. That is, this principle presupposes that you have well-defined credence in the proposition that you will be certain of what you learn.

with absolute certainty should find Subjective Evidential Trust quite plausible. Rationality recommends proportioning your credence to your evidence, not the negation thereof.

So the prospect of getting around Fact 2 by denying Evidential Certainty seems bleak. As long as the CA theorist accepts Weak Reflection, she must concede that, in some cases, one can be permitted and even required to plan to violate CA. But intuitively it should never be rational to plan to be irrational. So CA can't be a universal rational norm; it applies to some cases, but not others.

Let me turn now to Weak Reflection.

5.2 Weak Reflection

Recall that Weak Reflection says: if you know that you will update rationally on a genuine, transparent learning experience, then your prior should equal your prior expectation of your (planned) posterior:

$$\forall A \in \mathcal{F}_C, C(A) = \mathbb{E}_C[\tilde{f}(A)]$$

where $\tilde{f}$ is the map $w \mapsto f(\mathbf{E}(w))(A)$.

We can unpack this principle into the following two sub-claims:

1. **Disintegrability.** Your prior should equal your prior expectation of your planned future posterior, *if it were well-defined*, i.e.

$$\forall A \in \mathcal{F}_C, C(A) = \mathbb{E}_{\tilde{C}}[\tilde{f}(A)] = \int \tilde{f}(w)(A) d\tilde{C}$$

where $\tilde{C}$ is some probability measure (possibly finitely additive) over $(\Omega, \mathcal{P}(\Omega))$ that extends C .¹³

2. **Forethoughtfulness.** Your prior expectation of your planned future posterior should be well-defined. Equivalently: for every real number r and $A \in \mathcal{F}_C$, you should have well-defined credence in whether your future posterior in A will be above r (provided that you update in accord with f), i.e.

$$\forall A \in \mathcal{F}, r \in \mathbb{R}, A_r = \{w \in \Omega : f(A|\mathbf{E}(w)) \geq r\} \in \mathcal{F}_C.$$

A tempting thought is to reject Disintegrability. As we saw in Chapter 1, the CA theorist can't endorse Disintegrability in its full generality. And, the thought goes, if she has to draw a line somewhere, she might as well draw it at *Rational Difference*.

Unfortunately, it turns out that giving up Disintegrability alone is not enough to make CA compatible with Evidential Certainty, given Closure:

¹³Note: In Chapter 1, I took Disintegrability to be a constraint on rational conditional credences. Here I am treating it as a constraint on the relation between rational priors and updating plans.

Fact 4 (Blackwell and Dubins 1975). *If C is the Lebesgue measure over $[0, 1]$, then an updating plan f given C and $\mathcal{E}$ satisfies Evidential Certainty and Forethoughtfulness only if $f(E)$ is not countably additive for some (uncountably many) $E \in \mathcal{E}$.*

So the only option left for the CA theorist (apart from denying the existence assumption) is to reject Forethoughtfulness. Indeed, while CA is incompatible with Evidential Certainty and Weak Reflection in *Rational Difference*, it is compatible with Evidential Certainty and Disintegrability in that case.

Fact 5 (Berti and Rigo 1999, Theorem 2). *If C is the Lebesgue measure over $[0, 1]$, then there is an updating plan f given C and $\mathcal{E}$ that satisfies Evidential Certainty and Disintegrability and $f(E)$ is countably additive for every $E \in \mathcal{E}$.*

At first glance, this appears to be good news for CA. As noted in Chapter 1, Forethoughtfulness is an additional constraint only if the relevant partition is uncountably infinite (which is the case for *Rational Difference*). In such cases, one might think that rationality should allow us to cut some slack. If you have a well-defined prior credence in A , then it may be a rational requirement that you have well-defined posterior credence in A after you have updated on your evidence E , but you may not always be required to have well-defined estimates of your posterior credences in A , before your evidence comes in. You may cross the bridge when you come to it.

I have two reservations about this response. First, on some definition of credences, an agent have well-defined credence in a proposition just in case she can *grasp* that proposition, in some intuitive sense of grasping. Suppose this is true. Then one might think Forethoughtfulness follows from what it *means* to be an updating plan. To plan about what credence function you will adopt requires, in part, you being able to grasp the proposition that you will assign at least credence r to A (provided that A is a proposition you can already grasp). Easwaran (2013a) appears to express sympathy with this line of thought when he writes:

In planning whether to have credence greater than r , the agent is responding to whether the actual world is in A_r or not, and it seems that this responsiveness should require that A_r be a proposition she can grasp. (p.125, notations adapted)

In short, planning requires cognitive sophistication to process outcomes of the plan.

Second, since Fact 4 is a result of incompatibility, it shows that, assuming Closure and Evidential Certainty, CA rules out not only Forethoughtfulness but also the following weaker principle:

Weak Forethoughtfulness. If your (unconditional) prior and updating plan are both rational, then it is rationally *permissible* that your expectation of your planned posterior be well-defined.

But Weak Forethoughtfulness seems quite plausible. Even if one is not required to have well-defined expectations of one's own future credences, one should not be rationally prohibited from forming such expectations either. But it turns out that this must be true, if (i) it is rationally permissible for you to adopt the uniform Lebesgue measure as your unconditional prior in *Rational Difference*, and (ii) your updating plan in that case should preserve countable additivity and satisfy Evidential Certainty.

To take stock: Fact 1 shows that, assuming Conditionalization and Closure, CA is incompatible with Weak Regularity; Fact 2 shows that, assuming Closure and the much weaker principle Evidential Certainty, CA is incompatible with Weak Reflection (even if we relativize updating plan/conditional credence to partitions). Moreover, the incompatibility persists, even after we weaken Evidential Certainty to Subjective Evidential Trust, or Weak Reflection to Weak Forethoughtfulness.

Before proceeding, let me note an alternative interpretation of my two arguments. One way of reading Fact 1 and 2 (as well as the subsequent results) is that they present, not an objection, but a challenge to the CA theorist. For they show that, as one entertains more and more complicated learning scenarios, updating plans that always preserve countable additivity become harder and harder to come by. For the CA theorist, the principled response to this existential challenge would be to reject various principles that might otherwise seem intuitively plausible (e.g. Weak Regularity, the Multiplicative Axiom, Subjective Evidential Trust, Forethoughtfulness). But unless she replaces them with alternative constraints, it would appear that she is committed to a form of radical permissivism when it comes to updating on zero-probability events – practically, anything goes, as long as one satisfies CA. But this appears to be a significant cost of the view. Intuitively, even if we accept countable additivity as a lower bound on rational updating plans – an updating plan is irrational if it recommends credence functions that are not countably additive – it can't be an upper bound, even if the propositions that you will update on seem improbable by your own lights *ex ante*.

6 A way out

As noted at the beginning, both of my arguments rest on an existential assumption: in cases like *Random Point* and *Rational Difference*, there is a rational updating plan that specifies what credence function you should adopt given Levita's testimony. One might think this is where the CA theorist should get off the boat. After all, the cases that I have been looking at may seem somewhat fantastical. How could one draw a precise number between $[0, 1)$? And even if this can be done, how could they tell you that the number they picked belongs to the set of numbers that differ from it by a rational, without revealing to you strictly more information? For instance, what should Levita tell you, if she picked the number $\pi - 3$? And how will you process that information, within your limited life span?

Ultimately, I think this may be the most plausible response to my arguments. As we said before, infinities are weird, and some are weirder than others. However, this response also appears a last resort, *especially* for proponents of CA. For the whole point of CA is that credence functions with infinite domains are subject to constraints stronger than FA; infinitary credence functions do not lie beyond the reach of *synchronic* rationality. So, *mutatis mutandis*, it would appear that the CA theorist can't dismiss my arguments on the basis that infinitary learning situations nevertheless lie beyond the reach of *diachronic* rationality. Instead, they must think that only some of them do, while some of them do not. The challenge then again is to draw a line between those that do and those that don't on principled grounds that do not beg the question against her opponents. I am skeptical that this can be done. But even if there is such a line, it would mean that the paradise of countable additivity is much smaller than one may have expected.

Chapter 3

Borel and Bertrand

This chapter takes up a question left unresolved in our previous investigation: How should rational agents update on propositions to which they assign zero prior credence? In particular, is it always permissible to have well-defined conditional credence given zero-credence propositions? I look at one class of cases that appear to recommend a negative answer to this question: *the Borel-Kolmogorov paradox*.¹ One popular diagnosis of the paradox is that it shows conditional probability in general, and rational conditional credence in particular, to be intrinsically relativistic. When $P(E) = 0$, the value of $P(H|E)$ is only well-defined relative to the choice of a background partition that contains E as a member (or, in the more general setting, a sigma-algebra that contains E as an atom). As Kolmogorov himself put it, “the concept of a conditional probability with regard to an isolated given hypothesis whose probability equals 0 is inadmissible.”

This chapter has two parts. The negative part raises some problems for the Kolmogorov-inspired relativistic conception of rational conditional credence. The positive part proposes an alternative diagnosis of the paradox. Instead of treating it as ground for the radical claim that rational conditional credence must be gappy or relativistic, I contend that the paradox is an instance of a more familiar problem: the problem of priors. Sometimes, it is not obvious what rational credal attitude one should adopt in an epistemic situation. Whether there are zero, exactly one, or multiple credal attitudes justified in such situations is a separate question about which the paradox does not rule one way or another.

¹A historical note: Despite its name, a version of the Borel-Kolmogorov paradox was first introduced in Bertrand (1889). Admittedly, the question that Bertrand posed was slightly different. He asked: What is the probability that the distance between two randomly selected points X and Y on a uniform sphere is d ? Bertrand considered two approaches, both involve holding fixed a particular value of X . First, one might calculate the probability as the proportion of points that are d away from X . Second, one could calculate the probability as the proportion of points that are d away from X *on a particular great circle*. These two methods yield different answers. While many (most notably Borel) took the first method to be correct, they disagree about how the probability of the second method should even be calculated. That is, fixing a point X and a great circle passing through X , what is the probability that another randomly chosen point would be of certain distance away from X ? It turns out, as we show below, different methods for calculating *that* probability in turn disagree. This is what is now known as the Borel-Kolmogorov paradox (cf. Shafer and Vovk 2006).

Here is the plan: §1 explains the Borel-Kolmogorov paradox in more detail. §2 states the relativistic conception and discusses two objections to the view. §3 motivates my alternative diagnosis of the paradox. §4 concludes with some general philosophical reflections.

1 The paradox

Projectile. Two projectiles X and Y are launched toward an obstacle simultaneously. Let v_X, v_Y denote their respective speeds right before impact. Suppose you know that the chance distributions of v_X and v_Y are independent and uniform between 0m/s and 1m/s (exclusive). Let FAST denote the event that X hit the target at a speed greater than 1/2m/s and SAME SPEED the event that the two projectiles hit the target at the same speed. What credence should you assign to FAST, given SAME SPEED?

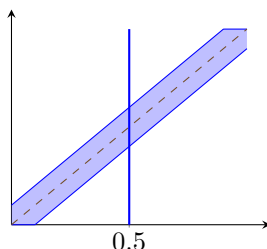
One might have thought that the answer should be 1/2. In Meehan and Zhang (2020), we considered two arguments in support of this judgment, and found both of them wanting. Without going into the fine details, let me note one reason for skepticism. Let CLOSE SPEED denote the event that the pre-impact speeds of the two projectiles are extremely close to each other, i.e. $|v_X - v_Y| \leq \epsilon$ for some specified ϵ that is small (e.g. $\epsilon = 0.000001$). One way of rationalizing our 1/2-intuition is by appealing to the following principle:

Speed Continuity. Your conditional credence in FAST given SAME SPEED should not be too far apart from your conditional credence in FAST given CLOSE SPEED, i.e.

$$C(\text{FAST}|\text{SAME SPEED}) \approx C(\text{FAST}|\text{CLOSE SPEED})$$

where $a \approx b$ just in case $|a - b| < \delta$ for some small enough δ .

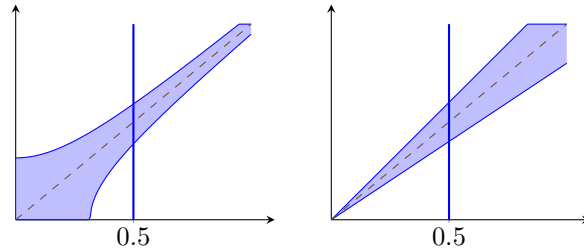
It follows from Speed Continuity that your conditional credence in FAST given SAME SPEED should be at least close to 1/2 (see the figure below for a pictorial proof). The principle also seems intuitively plausible. After all, it would be weird if rationality says, when you already know that the speeds of the two projectiles differ by at most 0.00000001m/s, learning that they are exactly equal to each other should make you significantly more or less confident in FAST. How could the elimination of such a small margin of error tell you anything substantive about the speed of X ?



The problem with Speed Continuity is that there are alternative measures of “closeness” that yield different, incompatible results. For instance, consider the following two metrics:

- Square Difference: v_X and v_Y are *close* just in case $|v_X^2 - v_Y^2| < \epsilon$ for some small ϵ ;
- Log Difference: v_X and v_Y are *close* just in case $|\log v_X - \log v_Y| < \epsilon$ for some small ϵ (equivalently, they are log-close if $|v_X/v_Y - 1| < \epsilon$).

It turns out that, if we measure “closeness” by the metric of square difference, then Speed Continuity entails that your conditional credence in FAST given SAME SPEED should be smaller than $1/2$. On the other hand, if we measure “closeness” by the metric of log difference, then Speed Continuity entails that your conditional credence in FAST given SAME SPEED should be greater than $1/2$.



More generally, say v_X is f -close to v_Y just in case $|f(x) - f(y)| < \epsilon$, where f is positive and strictly monotonic. Then, for every $r \in [0, 1]$, there is an f such that, if we define “closeness” by “ f -closeness,” then Speed Continuity entails that your conditional credence in FAST given SAME SPEED should be close to r (Arnold and Roberston 2019).

One may protest: Clearly we should measure closeness using absolute difference (i.e. the Euclidean metric). We *can* use alternative metrics, but the Euclidean metric is the most natural one, and it yields a result that coheres with our pre-theoretical judgment. So the burden of proof is still on the skeptic to show that the alternative metrics are real competitors to this default option.

Some (myself included) would find “naturalness” as an objective property deeply mysterious. It is also unclear that, even if there is one natural measure of closeness, it is the Euclidean metric as opposed to any other monotonic transformation thereof. But even if we grant the complaint, as noted in Meehan and Zhang (2020), the same point can be made by appealing to different choices of *units* or *quantities*, rather than metrics. There we discussed an imaginary case involving people who measure distance in “logameters” rather than “meters.” Here is a more mundane way to illustrate the same idea. Suppose both projectiles have the same mass of 2kg. Let E_X and E_Y denote their pre-impact (Newtonian) kinetic energies respectively. (Recall that $E = mv^2/2$. So numerically $E_X = v_X^2$ and $E_Y = v_Y^2$. Here we assume that you are rationally certain of the laws of Newtonian physics.) Again, intuitively, if you already know that E_X and E_Y differ by

at most $\leq 0.00000001J$, then the information that they are exactly equal to each other does not seem to give you substantive information about the speed of X . In other words, the following principle appears equally plausible:

Energy Continuity. Your conditional credence in FAST given SAME ENERGY should not be too far apart from your conditional credence in FAST given CLOSE ENERGY, i.e.

$$C(\text{FAST}|\text{SAME ENERGY}) \approx C(\text{FAST}|\text{CLOSE ENERGY})$$

where SAME ENERGY and CLOSE ENERGY denote respectively the events that $E_X = E_Y$ and $|E_X - E_Y| \leq \epsilon$ for some ϵ sufficiently small.

However, E_X and E_Y are close in absolute difference just in case v_X and v_Y are close in square difference. So by Energy Continuity, your conditional credence in FAST given SAME ENERGY should be much smaller than $1/2$.²

On the other hand, by assumption, you are rationally certain that $E_X = E_Y$ just in case $v_X = v_Y$. So presumably, your conditional credence in FAST given SAME ENERGY should equal your conditional credence in FAST given SAME SPEED. More generally, the following principle seems *prima facie* plausible:

Evidential Equivalence. If you are rationally certain that E_1 is necessarily equivalent to E_2 , then for any proposition H , it is a rational requirement that you assign the same credence to H given E_1 vis-a-vis E_2 .

But Speed Continuity and Energy Continuity are incompatible given Evidential Equivalence – a quantity cannot be both close to $1/2$ and much smaller than $1/2$. Notably, we arrived at this contradiction, *while holding fixed the background measure of closeness to be the Euclidean metric*.

Let's take stock: In *Projectile*, the following three claims seem individually plausible but are jointly incompatible:

1. (Speed Continuity) $C(\text{FAST}|\text{SAME SPEED}) \approx C(\text{FAST}|\text{CLOSE SPEED})$
2. (Energy Continuity) $C(\text{FAST}|\text{SAME ENERGY}) \approx C(\text{FAST}|\text{CLOSE ENERGY})$
3. (Evidential Equivalence) $C(\text{FAST}|\text{SAME SPEED}) = C(\text{FAST}|\text{SAME ENERGY})$.

One must give. But which one?

²The evaluation of the definite limit turns out to be quite mathematically involved. Just to give the reader a sense: For $\epsilon = 0.0001$, $C(\text{FAST}|\text{CLOSE ENERGY}) \approx 0.12$; for $\epsilon = 0.00001$, $C(\text{FAST}|\text{CLOSE ENERGY}) \approx 0.085$.

2 Relativism

As noted above, Kolmogorov concluded from the paradox that the concept of conditional probability given a zero-probability proposition is “inadmissible.” When $P(E) = 0$, the conditional probability of H given E is well-defined only with respect to the specification of a partition $\mathcal{E}$ that contains E as a member (or the specification of a sigma-algebra that contains E as an atom).³ Different choices of partitions may yield different conditional probabilities of H given E . In such cases, there is no such thing as “the” conditional probability of H given E , just as there is no “the” limit of $f(x)$ at a point of discontinuity. This is not a paradox, but a fact about conditional probability that one has to live with.

One way of interpreting Kolmogorov’s proposal in the epistemic setting is that rational conditional credence should be represented by a three-place function rather than a two-place function (cf. Easwaran 2008). Equivalently, we should model an agent’s credal attitudes by what I call a **credal state** (cf. Chapter 1), which is a partial function that maps partitions to binary conditional credence functions subject to the constraint of agreement given the tautology. More precisely, a credal state is a function of the form

$$C : \mathcal{E} \mapsto C_{\mathcal{E}}(.|.) : \mathcal{P}(\Omega) \times \mathcal{E} \cup \{\Omega\} \rightarrow \mathbb{R}$$

with $C_{\mathcal{E}}(.|\Omega) = C_{\mathcal{E}'}(.|\Omega)$ (whenever the two are well-defined.) While rationality requires that each $C_{\mathcal{E}}$ be coherent in the sense of Definition 1, it does not impose any matching constraints amongst the $C_{\mathcal{E}}$ ’s. In particular, for some H and $E \in \mathcal{E} \cap \mathcal{E}'$, rationality may permit and even require that $C_{\mathcal{E}}(H|E) \neq C_{\mathcal{E}'}(H|E)$. In such cases, one’s conditional credence in H given E *should* be undefined, when the relevant partition is unspecified. Call this view *Relativism*.⁴

The Relativist gets around the inconsistent triad by denying Evidential Equivalence. The thought is: Even if you are rationally certain that SAME SPEED is necessarily equivalent to SAME ENERGY, you are required to adopt different conditional credence in FAST given SAME SPEED vis-a-vis SAME ENERGY, when the two events are viewed as members of two different partitions of the unit square. More specifically, let $\mathcal{E}$

³In fact this is not entirely correct: In Kolmogorov’s framework, conditional probability is a special case of conditional expectation, which is only well-defined up to measure-zero set. So, after selecting a partition/sigma-algebra $\mathcal{E}$, one must in addition choose a version of the conditional expectation given $\mathcal{E}$, in order to obtain a unique number that could be interpreted as $P(H|E)$. Sometimes there may be one “canonical” version based on continuity/symmetry-considerations. But whether such cases constitute the norm or the exception is a matter of perspective. Here I shall set aside this technical complication and speak as if picking a partition/sigma-algebra is sufficient for determining a unique value of the relevant conditional probability/rational conditional credence.

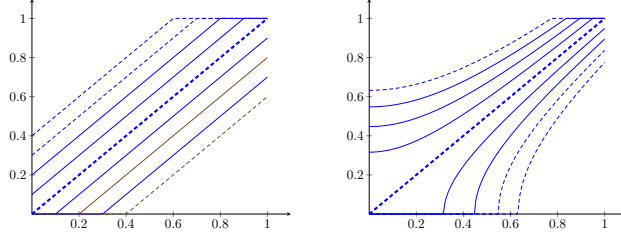
⁴Alternatively, we could model credal attitudes by partial functions from *partition-proposition pairs* to unconditional credence functions:

$$C : (E, \mathcal{E}) \mapsto C_{\mathcal{E}}(.|E) : \mathcal{P}(\Omega) \rightarrow \mathbb{R}.$$

This brings out a contrast between Relativism and the theory of inquisitive probability developed by Hoek (2020b) (Hoek calls pairs of the form $(E, \mathcal{E})$ “quizpositions”). One notable feature of Hoek’s theory is that any two intensionally equivalent quizpositions must receive the same inquisitive credence (insofar as one has opinions about them in the first place). While Hoek defines inquisitive credence as a unary function, a natural generalization of his theory would suggest that it respects Evidential Equivalence, which is what the Relativist denies.

and $\mathcal{E}'$ be defined as follows:

$$\mathcal{E} = \{(v_X, v_Y) : |v_X - v_Y| = r : r \in [0, 1]\} \quad \mathcal{E}' = \{(v_X, v_Y) : |E_X - E_Y| = r : r \in [0, 1]\}.$$



Speed Continuity constrains $C_{\mathcal{E}}(\text{FAST}|\text{SAME SPEED})$, whereas Energy Continuity constrains $C_{\mathcal{E}'}(\text{FAST}|\text{SAME ENERGY})$. But since $C_{\mathcal{E}}$ and $C_{\mathcal{E}'}$ are two different conditional credence functions, they do not have to agree with each other given the same event, i.e.

$$C_{\mathcal{E}}(\text{FAST}|\text{SAME SPEED}) \neq C_{\mathcal{E}'}(\text{FAST}|\text{SAME ENERGY}).$$

Singpurwalla and Swift (2001) summarize this approach nicely [notations adapted]:

[T]he cause of what is viewed as a paradox is the freedom (or flexibility) that is given to us in defining the diagonal and characterizing its vicinity. Each choice reflects a particular partition (discretization) of the sample space, and the geometry of each partition leads us to its own answer for the uncertainty of v_X .

3 Critique of Relativism

But one could press: *Why* should one's choice of a partition rationally influence one's uncertainty of v_X ? What normative difference could it make, whether you view SAME SPEED as a member of $\mathcal{E}$ or $\mathcal{E}'$?

The answer to this question depends on the representational content of the relevant partition. In Chapter 1, I suggested one proposal: Each partition specifies a *question*, or more broadly an *inquisitive context*. The thought is: One does not conditionalize on E *simpliciter*, but always E *qua* an answer to some specified question, whose possible answers form a partition of the sample space. For instance, in *Projectile*, the partitions $\mathcal{E}$ and $\mathcal{E}'$ correspond to the following two questions respectively:

Q1: What is the absolute difference between the speeds of the two projectiles when they hit the target?

Q2: What is the absolute difference between the kinetic energies of the two projectiles when they hit the target?

How confident one should be in FAST given SAME SPEED depends on whether SAME SPEED answers Q1 or Q2. If it answers Q1, then $C(\text{FAST}|\text{SAME SPEED}) = 0.5$; if it answers Q2, then $C(\text{FAST}|\text{SAME SPEED}) < 0.5$.

As noted in Chapter 1, this explanation is not supposed to be intuitive or compelling. To rehearse a worry made there: Suppose you ask a reliable source about the absolute difference between the speeds of the two projectiles right before impact, and they answer 0. You judge that you have learnt no information that should change your mind about FAST; you are still indifferent about whether the projectile X hit the target at a speed of at least $\frac{1}{2}\text{m/s}$. However, suppose the source comes back to you with an apology,

“Sorry I made a mistake – I thought you were asking about the absolute difference between their kinetic energies. But don’t get me wrong. What I said was true, just that it was for a different question.”

Should you change your mind in FAST, after the source’s retraction? It seems to me that you shouldn’t; but the Relativist must think that you should. The 0-answer is irrelevant to FAST as an answer to Q1, but it is evidence against FAST as an answer to Q2.

Is there a way of making the view more palatable? I am skeptical; but it would be pre-mature to jump to conclusions before we have explored all options. Here I shall examine two ways of justifying Relativism: *hyperintensionalism* and *contrastivism*. I argue that neither is fully satisfactory. While my objections are not meant to be definitive, taken together, I think they motivate the search for an alternative solution to the Borel-Kolmogorov paradox.

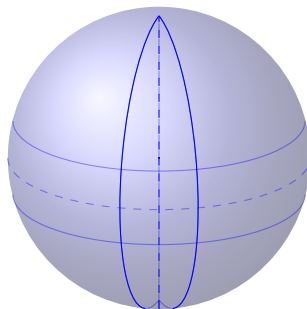
Before proceeding, let me mention two other solutions to the paradox that have been made in the literature, only to set them aside for now. My main concern for both is that they are not general enough; in particular, it is unclear how either helps us get around the inconsistent triad of *Projectile*. (This part can be skipped without loss of continuity.)

Some think the conditional probability of H given E may be undefined because E is imprecise. To motivate this idea, I need to introduce the original Borel-Kolmogorov paradox. Suppose a point is chosen at random on a uniform sphere. Consider the following two questions:

- (i) What is the conditional probability that the point has a latitude between 30 degrees north and 30 degrees south, given that it is on the Meridian;
- (ii) What is the conditional probability that the point has a longitude between 30 degrees east and 30 degrees west, given that it is on the Equator.

Naively, the answers to these two questions should be the same. After all, it is a mere convention whether we label a great circle as the Meridian or the Equator. However, the received wisdom is that they are not

the same. The first one is $1/2$, whereas the second one is $1/3$. Intuitively, this is because, although the Meridian and the Equator are both great circles, when computing their respective probability densities, one views them as “constructed” from two different limiting procedures: one from a sequence of circular bands of uniform widths, the other from a sequence of opposing wedges connected at the two poles (see Figure below). As a result, points on the Equator are uniformly distributed, whereas points on the Meridian are concentrated towards the Equator and away from the poles.



Jaynes (2003) gives the following diagnosis of the paradox:

Nearly everybody feels that he knows perfectly well what a great circle is; so it is difficult to get people to see that the term ‘great circle’ is ambiguous until we specify what limiting operation is to produce it. The intuitive symmetry argument presupposes unconsciously the equatorial limit; yet one eating slices of an orange might presuppose the other. (p.470)

So according to Jaynes, the paradox arises from an illusion of precision. We think there is one definite object that is “the great circle.” But really there are many “great circles,” each arising from a distinct limiting procedure. The limit of a band is not the same as the limit of an eating slice of orange. While we can define an abstract object that is the equivalence class of these different limiting constructions, and call it “the great circle,” their differences turn out to matter in the probabilistic context. So we need to relativize conditional probability to the choice of a partition or a sigma-algebra, which in turn (supposedly) specifies one unique limiting construction.

My main problem with Jaynes’ analysis as a general solution to the Borel-Kolmogorov paradox is that it does not seem to apply to cases like *Projectile*. In order to explain away the inconsistent triad in a similar fashion, one must say that SAME SPEED (or SAME ENERGY) is somehow imprecise because the notion of “equality” is ambiguous until one specifies what limiting operation is to produce it.⁵ This seems to me a

⁵Note that Jaynes is saying that the proposition *A randomly selected point lies on a great circle* is imprecise (because what a great circle is is under-specified). He is not saying that the proposition *A randomly selected point lies “close” to a great circle* is imprecise. Therefore, when applying Jaynes’ diagnosis to *Projectile*, the relevant imprecise proposition would have to be SAME SPEED rather than CLOSE SPEED. (Also recall that CLOSE SPEED stands for the proposition that the absolute difference between the speeds of the two projectiles is smaller than ϵ for a fixed ϵ , and is therefore precise insofar as set-theoretic representation goes).

harder pill to swallow. If anything, it strikes me that “equality” (more precisely “numerical equality”) is one of the most unambiguous mathematical notions on which we build everything else.⁶

A related idea is that rational conditional credence should be relativized to the choice of a partition because propositional attitudes are in general sensitive to representational perspectives. For instance, many think mathematical facts are true necessarily. So $1234567 = 1234567$ just in case $1234567 = 127 \times 9721$. On the other hand, while it may be a rational requirement that

$$C(1234567 \text{ is a prime} | 1234567 = 127 \times 9721) = 0$$

it seems rationally permissible to be uncertain whether 1234567 is a prime given that it is self-identical, i.e. it is rational that

$$C(1234567 \text{ is a prime} | 1234567 = 1234567) > 0.$$

For another example, most philosophers think Hesperus is necessarily identical to Phosphorus. In particular, Hesperus has craters just in case Phosphorus has craters. But it seems rational for an ancient Greek astronomer to be fully confident that Hesperus has craters given that it has craters, but be uncertain whether Hesperus has craters given that Phosphorus has craters.

These cases show that, in general, one may believe/assign certain credence to p , but fail to believe/assign the same credence to its necessarily equivalent represented in a different way. A tempting thought is that the Borel-Kolmogorov paradox is an instance of the same phenomenon. Just as there is nothing paradoxical in the rational combination of

$$C(1234567 \text{ is a prime} | 1234567 = 127 \times 9721) = 0$$

$$C(1234567 \text{ is a prime} | 1234567 = 1234567) > 0$$

so too there is nothing paradoxical in the rational combination of

$$C(\text{FAST} | \text{SAME SPEED}) = 0.5$$

$$C(\text{FAST} | \text{SAME ENERGY}) < 0.5.$$

As Rescorla put it,

Credences are rationally sensitive to the way that one represents an event, so it is not surprising

⁶A related proposal that I have come across in conversation (though not in print) is that the choice of a partition/sigma-algebra signifies a measurement procedure. However, it is unclear to me why the way in which a quantity is measured should make a rational difference, once one knows its value (see below).

that different conditional probabilities arise when one represents the conditioning event differently. The Borel-Kolmogorov paradox is a Fregean “paradox of identity.”

However, there seems to be one important disanalogy. It appears that rational credence is sensitive to ways in which one represents an event, *only if* the agent is ignorant that they are different ways of representing the same event. It would be irrational to be uncertain whether 1234567 is a prime, if one *knows* that it is the product of 127 and 9721. Similarly, it would be irrational for the ancient Greek astronomer to be uncertain whether Hesperus has craters, if he knew that Phosphorus has craters and Hesperus is Phosphorus. On the other hand, in *Projectile*, the combination of

$$C(\text{FAST}|\text{SAME SPEED}) = 0.5$$

$$C(\text{FAST}|\text{SAME ENERGY}) < 0.5.$$

appears rational (insofar as Speed Continuity and Energy Continuity are individually plausible), even when the agent knows that SAME SPEED is necessarily equivalent to SAME ENERGY.⁷

I take neither consideration to be definitive. One might also think there are multiple “Borel-Kolmogorov paradoxes” that call for different analyses. I am partially sympathetic to this proposal. However, for the time being, I shall proceed under the assumption that there is one unified paradoxical phenomenon, of which *Projectile* is a paradigm example.

3.1 Hyperintensionalism

One way of justifying the question-sensitivity of rational conditional credence is to understand a *question* as specifying a *subject matter*. The thought is: What rational conditional credence one should adopt given E depends on what E is *about*. While SAME SPEED and SAME ENERGY are equivalent given the Newtonian laws of physics, they are inequivalent in terms of their subject matters; one is about difference in speed, the other one is about difference in kinetic energy.

An analogue of this idea has been explored in the context of confirmation. Pre-theoretically, one might have the intuition that the observation of a black raven confirms the hypothesis *All ravens are black*, but not the hypothesis *All non-black things are non-ravens*, even though the two are (classically) logically equivalent to each other. The right response to the Raven paradox remains an issue of contention (e.g. Fitelson and

⁷Rescorla concedes this point. He agrees that treating the Borel-Kolmogorov paradox as a Frege puzzle does not address the problem of rational credal allocation when the agent tries to “coordinate distinct representational perspectives with one another” and considers this to be “an undeniable limitation of Kolmogorov’s framework.” (p.761) To this extent I think we only disagree over the terminological issue of what the Borel-Kolmogorov paradox is: I take the paradox to arise only when the agent is rationally certain of the equivalence amongst the conditioning propositions, whereas (it appears that) Rescorla takes the paradox to have a wider scope.

Hawthorne 2010; Rinard 2014). One proposal that has been put forward is that the confirmation relation tracks differences in subject matter (Goddard 1977; Yablo 2014). In fact, Hempel himself toyed with this idea (only to reject it thereafter):

Perhaps the impression of the paradoxical character of [these cases] may be said to grow out of the feeling that the hypothesis that *All ravens are black* is about ravens, and not about non-black things, nor about all things (1945, p.17).

Indeed, if one adopts the probabilistic account of confirmation, i.e. E confirms H relative to a rational credence function C just in case $C(H|E) > C(H)$, then every Borel-Kolmogorov paradox can be re-construed as a paradox about confirmation. For instance, we can reformulate the inconsistent triad of *Projectile* as follows:

1. (Speed Continuity*) SAME SPEED confirms FAST to approximately the same degree as CLOSE SPEED.
2. (Energy Continuity*) SAME ENERGY confirms FAST to approximately the same degree as CLOSE ENERGY.
3. (Evidential Equivalence*) SAME SPEED confirms FAST just in case SAME ENERGY confirms FAST.

The current proposal is that one should give up Evidential Equivalence*: SAME ENERGY disconfirms FAST but SAME SPEED does not, because they are about different things. More generally, whether E confirms H /raises its probability depends on the subject matter of both H and E .

I have two concerns about this proposal. First, it is unclear to me that differences in subject matter should make a rational impact on one's conditional credence. Consider the following variant of *Projectile*:

Projectile Redux. Same set-up as before, except that this time you have two measurement devices, A and B. A measures the absolute difference between the speeds of the two projectiles accurately and with infinite precision; B measures the absolute difference between the kinetic energies of the two projectiles, also accurately and with infinite precision. Note that A reads 0 if and only if B reads 0, and you know this *ex ante*. The two projectiles hit the target. You learn that one device reads 0.

Does it matter which device that reading came from? Intuitively, the answer seems to be No. After all, no matter where it came from, you *know* the reading of the other device – it must be 0 as well. So there seems to be no point in resolving that uncertainty. But on the hyperintensionalist approach, there is – if the reading came from device A, then it is *about* the difference in speeds, whereas if it came from device B, then it is *about* the difference in kinetic energies. In one case, you get no evidence about FAST; in the other case, you get evidence against FAST.

My second concern is more flat-footed: Intuitively, if $C(H|E)$ is sensitive to the subject matter of E , then it should be so whether or not $C(E) = 0$. But the Relativist concedes that, when $C(E) > 0$, $C(H|E)$ is well-defined by the ratio formula and therefore *insensitive* to the subject matter of E . This discontinuity seems puzzling. Why should difference in subject matters makes a normative impact, only when the proposition receives zero unconditional credence?

Shafer (1985) raises essentially this worry against Kolmogorov’s proposed solution. He points out that, even if one agrees with Kolmogorov that conditional probability given null proposition is inadmissible, “some puzzlement remains.” For “[i]f conditioning on an isolated event is meaningful in the discrete case, why is it not meaningful in the general case?”

Shafer also has an answer to this question:

The answer, of course, is that conditioning on an isolated event is not meaningful even in the discrete case. Conditioning is legitimate and meaningful only when there is a protocol which specifies what else might have been observed and hence determines a partition.

This is the contrastivist version of Relativism, which I shall explore in the next subsection. At this point, however, it is worth emphasizing that the problem of discontinuity is a problem that every solution to the Borel-Kolmogorov paradox must wrestle with. For fundamentally it concerns a tension between two intuitive thoughts. On the one hand, the paradox arises only with respect to conditionalizing on zero-credence propositions, so such cases must be special. On the other hand, there seems to be nothing normatively special about conditionalizing on zero-credence propositions. Our best theory of rationality should not make a sharp distinction between how one should update on a zero-credence (contingent) proposition versus how one should update on a 0.000001-credence proposition. It appears that the hyperintensionalist must embrace that discontinuity. The question is whether there are alternative versions of Relativism that can do better. This brings us to *contrastivism*.

3.2 Contrastivism

On a first approximation, *contrastivism* is the view that how one should update on a piece of information depends on what one could have learnt instead.⁸ Rational conditional credence is sensitive to the choice of a partition because conditionalization on E is illegitimate unless one specifies what else might have been learnt. Or, to put the point slightly differently: One does not conditionalize on E *simpliciter*, but always E *as opposed to* some other salient propositions that are mutually exclusive and (with E) jointly exhaustive.

⁸I borrow this formulation of “contrastivism” from Gallow (2019), who discusses a similar principle with respect to evidential import.

But one might press the question: Why should rationality require sensitivity to information that one does not have, but could have had? You could have seen an UFO when you look outside of your window; you could have been reading a paper on the feeding behavior of red pandas instead of probability paradoxes (which may have been a better use of your time anyway). But those possibilities seem neither here nor there, with respect to what you should actually believe right now given your actual evidence.

Contrastivists often draw inspiration from the following classic probability puzzle (e.g. Billingsley (2008) and Easwaran (2019)):

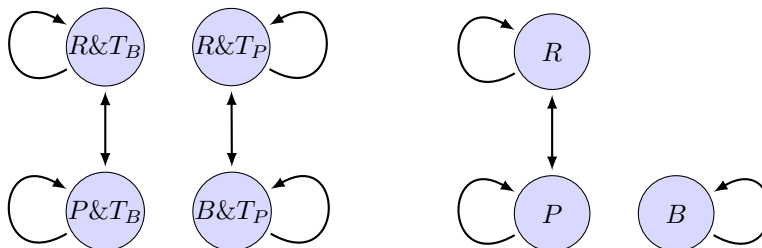
Three Prisoners. Percy, Bill and Roy are being tried for espionage. Two of them will be sentenced to life in prison; one will be set free. The person who will be pardoned is determined by a fair lottery, and only the prison guard knows who it is.

As they wait for the judgment day, Roy comes up with an idea:

“Right now I am $1/3$ confident that I will be set free. However, I could ask the guard to name one person who will be condemned, other than myself. Whatever he tells me, it means that the person who will be pardoned is either me or the other person who is not named. So I’ve raised my odds of freedom from $1/3$ to $1/2$ – how smart!”

Roy’s reasoning seems fallacious. But where’s the mistake? One popular diagnosis is that Roy is irrational because he has failed to take into account what the guard could have told him otherwise. Specifically, since the guard can only name one name, he must make a choice when it is Roy who will be set free. As a result, given the guard’s testimony, the probability that Roy will be set free is going to be lower than the probability that the other person will be set free.

Contrast this with an alternative scenario: Suppose the guard will tell Roy whether Bill will be condemned. He won’t mention a word about Percy. In that case, it would be rational for Roy to be $1/2$ confident that he will be set free, when he learns that Bill will be condemned. But that’s because, if he learns that Bill will not be condemned, he knows he is not the lucky one. So given the same testimony, Roy should revise his beliefs differently, depending on what the guard could have told him otherwise.



Pearl (1988) draws the following conclusion from this example:

The lesson of the Three Prisoners paradox is that we cannot assess the impact of new information by considering only propositions implied by the information; we must also consider what information *could have* been reported. (original emphasis)

So we seem to have an abductive argument for contrastivism; the view is true because it offers the best explanation of what went wrong in cases like *Three Prisoners*.

One problem with this argument is that there seems to be an alternative, better explanation available: Roy is irrational, not because he has failed to take into account what he could have learnt, but because he has not taken into account everything that he has learnt. In particular, when the guard tells him that, e.g. Percy will be condemned, he has learnt something stronger, namely that *the guard told him that Percy will be condemned*. It is straightforward to show that, given this stronger proposition, Roy should be 1/2 confident that he will be set free only if Roy is confident that the guard resolutely avoids interacting with Bill unless necessary.⁹ But in that case, Roy should be confident that he will not be set free, if the guard had given the letter to Bill instead. Since the cases of Percy and Bill are symmetrical, it follows that Roy should not be 1/2 confident that he will be set free irrespective of the guard's testimony. In short, *Three Prisoners* supports not contrastivism, but the principle of total evidence (PTE).

Now, most theorists who see a parallel between *Three Prisoners* and the Borel-Kolmogorov paradox don't disagree with the above diagnosis. What this shows, they argue, is that contrastivism is implicit in PTE.¹⁰ To update on your total evidence invariably involves being sensitive to what you could have learnt instead.

But this seems puzzling. For on a naive reading, PTE is *inconsistent* with contrastivism; the former says that one should take into account all and only one's actual total evidence; the latter says that one should take into account what one could have learnt as well.

⁹Let T_P be the proposition that the guard tells Roy that they have given the letter to Percy. Then

$$P(R|T_P) + P(B|T_P) = 1.$$

So $P(B|T_P) = 1/2$ if $P(R|T_P) = 1/2$. By Bayes' rule,

$$P(B|T_P) = \frac{P(T_P|B)P(B)}{P(T_P|B)P(B) + P(T_P|R)P(R)} = \frac{1}{1 + P(T_P|R)}.$$

So $P(B|T_P) = 1/2$ only if $P(T_P|R) = 1$, i.e. the guard will give the letter to Bill only if Percy will be released.

¹⁰For instance, Shafer writes,

The need for a protocol is implicit in the principle that you should condition on everything you learn. To see this, notice that when you learn something, you also learn that you have learned it...So the event B that specifies all the facts you learn must also specify the fact that you learn these facts....But when we say that you should condition on B, we are assuming that B is in your probability model...So it is implicit in the principle of total evidence that your probability model should include a model for what you learn. It should give probabilities for the different possible ways your learning may turn out. This amounts to giving a protocol [i.e. partition] for new information. (p.264)

However, it is unclear to me why the last step should follow, i.e. why assigning probabilities to different possible courses of learning experience "amounts to" giving a protocol/specifying a partition. For presumably a non-Relativist may readily concede that one assigns probabilities to many partitions, while denying that those partitions are relevant for determining one's rational conditional credence/posterior credence in H given E .

I think this confusion brings out the need to distinguish between two contrastivist theses, which in turn calls for a distinction between the notion of *total evidence* and what I'll call *direct information*. Suppose you are about to undergo a learning experience, e.g. you are about to open the *New York Times* website to check the latest update on COVID statistics. Suppose the website says that the number of new cases has dropped by 50% in the last two weeks. The proposition that *the number of new cases has dropped by 50% in the last two weeks* – call it p – is now part of your *direct information*. It is a piece of information that is communicated directly to you by the source.

By contrast, consider the following proposition:

Lp : You have learnt that the number of new cases has dropped by 50% in the last two weeks.

Unlike p , Lp is not going to be something that you will read on the website. Nevertheless, suppose all goes well, when you see the data, you will learn Lp in addition to p . For instance, you may report to your friend, “I have just learnt some good news – the number of new cases of COVID has dropped by 50% in the last two weeks.” The evidential basis of your assertion is Lp , rather than p .

Similarly, in *Three Prisoners*, when the guard tells Roy that Percy will be condemned, Roy acquires the direct information that *Percy will be condemned*. But he also learns something else, namely that *he has learnt from the guard that Percy will be condemned*. The former is about Percy; the latter is about Roy.

So, if we identify total (incremental) evidence with the strongest, true proposition that one learns in a learning episode, then direct information may not exhaust one's total evidence. In fact, it may not even be part of one's total evidence. You may be hallucinating; the guard may be lying.

We can now distinguish between two different contrastivist theses:

Direct Information Contrastivism (DIC). In a learning episode, how one should update on a piece of direct information depends on (one's evidence of) what direct information one could have obtained instead.

Total Evidence Contrastivism (TEC). In a learning episode, how one should update on one's actual total evidence, depends on (one's evidence of) what total evidence one could have learnt instead.¹¹

The contrastivist is right that DIC follows from the principle of total evidence. For according to that principle, rational agents should update on their total evidence, and cases like *Three Prisoners* suggest that whether a piece of direct information (e.g. *Percy will be condemned*) constitutes one's total evidence depends on what other direct information one could have obtained instead.

¹¹The parenthetical remarks are there because some philosophers (e.g. evidentialists with an internalist bent) might think that rationality only requires sensitive to facts that are internally accessible to the agent, given their total evidence. According to this view, DIC is true only insofar as two rational agents may update differently in response to the same piece of direct information because they have different evidence about what alternative direct information they could have obtained instead. While I am attracted to externalism, the difference between these two versions of DIC and TEC do not matter much for the purposes of our present discussion.

However, the principle that we would need in order to vindicate Relativism is not DIC but TEC. For by stipulation, in *Projectile*, your learning that the two projectiles have the same speed/kinetic energy is both necessary and sufficient for them having the same speed/kinetic energy; you would have learnt something different if and only if their speeds/kinetic energies were different. (By contrast, in *Three Prisoners*, Percy may be condemned even if the guard does not tell you so.) So SAME SPEED will be your total evidence if it is true. And the Relativist thinks, even in that case, how you should update on that piece of information depends on what you could have learnt otherwise.¹²

To take stock: One argument for contrastivism is that it is the best explanation for the irrationality of Roy's reasoning in *Three Prisoners*. I have argued that there is an alternative, better explanation: Roy is irrational because he violates the principle of total evidence. While PTE is compatible with Direct Information Contrastivism, it is incompatible with Total Evidence Contrastivism, which is the version of contrastivism that we need in order to vindicate the Kolmogorovian treatment of the Borel-Kolmogorov paradox (in particular its rejection of Evidential Equivalence).

Now there is a wrinkle with my argument: My diagnosis of Roy's irrationality assumes that his learning experience is transparent, in the sense that if he learns E , then he also learns that he learns E . But this assumption is controversial. Some think not all learning experiences are transparent; one could learn E , but fail to learn that one has learnt E . To give a concrete example, consider the following variant of *Three Prisoners*:¹³

Opaque Three Prisoners. Same set-up as above, except that this time, the guard will tell Roy the identity of a person who will be condemned by telepathy instead of verbal communication. Moreover, a bad introspector that he is, Roy will acquire that information without realizing that he does.

So Roy's learning experience is non-transparent. When he learns that, say, Percy will be condemned, he does not automatically acquire the introspective knowledge that he has learnt that piece of information.

Now suppose Roy reasons to himself

Right now I am $1/3$ confident that I will be set free. But no matter whose name the guard will plant into my head, it will mean that the person who will be pardoned is between me and the other person who is not named. So I've raised my chance of freedom from $1/3$ to $1/2$ – how smart!

¹²To put the point slightly differently: In *Three Prisoners*, whether Roy should update by conditionalizing on his direct information depends on what he could have learnt instead (because he should conditionalize on his total evidence, and whether his direct information constitutes his total evidence depends on what he could have learnt instead). By contrast, in *Projectile*, the question is not *whether* but *how* you should conditionalize on your direct information (i.e. SAME SPEED), under the assumption that it is your total evidence.

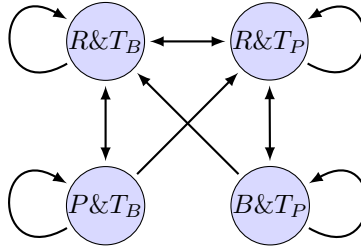
¹³Bronfman (2014) considers a version of *Opaque Three Prisoners*. Bronfman concludes that, insofar as these cases are possible, one should reject Bayesian conditionalization as the universally rational updating procedure. Others draw different conclusions. See the next chapter for further discussion.

At first glance, Roy's reasoning seems just as bad as his counterpart in *Three Prisoners*. However, this time, Roy has not violated the principle of total evidence. For by stipulation, no matter who is named, he will learn the name and nothing else. So his direct information would constitute his total evidence. By Conditionalization, irrespective of what the guard tells Roy by telepathy, Roy should be 1/2 confident that he will be pardoned, after the guard plants the name into his head.

Before proceeding, a quick technical aside. Note that in *Opaque Three Prisoners*, Roy's credences are conglomerable in the sense of Chapter 1. However, he is inconglomerable in a more general sense. Let C be an unconditional credence function and f_C an updating plan given C and a learning experience $\mathbf{E}$ (cf. Chapter 2). We say (C, f_C) is **E-conglomerable** if, for all $E \in \mathcal{E} = \{\mathbf{E}(w) : w \in \Omega\}$ and $A \in \mathcal{F}_C$,

$$\inf_{E \in \mathcal{E}} f_C(E)(A) \leq C(A) \leq \sup_{E \in \mathcal{E}} f_C(E)(A).$$

Then conglomerability as defined in Chapter 1 corresponds to conglomerability in this more general sense of the pair $(C, C(\cdot|\cdot))$ with respect to a partitional learning experience (i.e. $\mathcal{E}$ is a partition). On the other hand, in *Opaque Three Prisoners*, Roy's learning experience is non-partitional, and what we saw above is that the pair of uniform prior and conditionalization plan is inconglomerable (in the more general sense) with respect to that learning experience.



To the extent that Roy is irrational in both *Three Prisoners* and *Opaque Three Prisoners*, the principle of total evidence can't be the whole story. We need an alternative explanation, and contrastivism may seem a good fit. Just as, in the transparent case, how Roy should update on the guard's verbal testimony depends on what the guard could have told him otherwise, one might think that, in the non-transparent case, how Roy should update on the guard's telepathic message depends on what telepathic message the guard could have planted into his head instead. Whether that information is his total evidence or direct information is irrelevant. The crucial point is, as Pearl put it, that the impact of new information depends on what information could have been reported.

So, the thought goes, contrastivism and in particular TEC is true, because it provides the best explanation of what went wrong in cases like *Opaque Three Prisoners*. Rationality requires due sensitivity to actual as

well as counterfactual evidence. This is why one should deny Evidential Equivalence in *Projectile*: Like Roy in *Opaque Three Prisoners*, how you should update on SAME SPEED, when it is your actual total evidence, depends on what you could have learnt instead. If you could have learnt the absolute difference between the speeds of the two projectiles, then you should remain 1/2 confident in FAST given SAME SPEED; if you could have learnt the absolute difference between their kinetic energies, then you should be less than 1/2 confident in FAST given SAME SPEED.

One tempting response to this argument is to deny that the two Roys in *Three Prisoners* and *Opaque Three Prisoners* are rationally on a par. The Roy in *Three Prisoners* is irrational, but the Roy in *Opaque Three Prisoners* is rationally impeccable. After all, one moral from Chapter 1 is that we have to live with some form of rational inconglomerability, if we want to make room for higher orders of infinity in our epistemology. Perhaps the same is true with respect to non-partitionality.

I'll defer a fuller discussion of cases like *Opaque Three Prisoners* to the next chapter.¹⁴ But let me confess upfront that I am torn about this case. Roy's reasoning strikes me as irrational, and yet it is unclear how he could have reasoned differently. My preferred diagnosis, which I develop in more detail in Chapter 4, is that, insofar as cases like *Opaque Three Prisoners* are possible,¹⁵ they reveal a fundamental tension between two different evaluative perspectives on rational updating: *ex ante* and *ex post* (cf. Huttegger (2017)). *Ex ante* Roy should not *plan* to revise his credence in a predictable direction, no matter what he learns; *ex post* it is rational for Roy to *end up* revising his credences in a predictable direction, no matter what he learns. If this is right, then we have discovered a surprising fact about epistemic rationality: rational updating plans may not always be binding. Sometimes, it is rational to deviate from the rationally planned trajectory.

But we are getting ahead of ourselves. At present, I think there is a more flat-footed response available. Again, let's distinguish between two different contrastivist theses:

Total Evidence Contrastivism-1. In a learning episode, how one should update on one's actual total evidence depends on *whether one's possible total evidence forms a partition*.

Total Evidence Contrastivism-2. In a learning episode, how one should update on one's actual total evidence depends on *which partition corresponds to one's possible total evidence* (provided that it does form a partition).

¹⁴This is also related to a preliminary point I made in the Introduction. As noted there, the norm Conditionalization is ambiguous and admits of at least 32 different possible precisifications. One central parameter is whether the norm assumes that the agent's possible evidence forms a partition of her doxastic space. As I show in Chapter 4, one class of arguments for Conditionalization – the instrumentalist argument (which includes the Dutch book argument as a special case) – do not go through without the assumption of partitionality. How to make of this result remains a point of controversy. My own preferred approach is to hold onto the core instrumentalist commitment (i.e. rational updating plans are instrumentally optimal) but adopt a different formal measure of instrumental optimality. I explain this approach in more detail towards the end of the next chapter.

¹⁵Some take this tension as evidence that cases like *Opaque Three Prisoners* are metaphysically impossible; while I am tempted by that thought, I am also skeptical of drawing metaphysical conclusions from normative conflicts.

The two contrastivist theses are logically independent of each other. In particular, TEC-1 does not entail TEC-2; it is conceivable that rational updating is sensitive to whether one’s possible evidence forms a partition, but insensitive to which partition it forms. The key observation is that, while *Opaque Three Prisoners* may support TEC-1, it does not support TEC-2. In fact it is irrelevant to TEC-2, since Roy’s learning experience is non-partitional by stipulation.

On the other hand, what we need in order to vindicate the Relativist verdict about *Projectile* is TEC-2 rather than TEC-1. Again, by stipulation, your learning experience in *Projectile* is partitional. You will either learn the precise difference between the speeds of the two projectiles, or the precise difference between their kinetic energies. In each case, no pair of information that you might obtain may overlap with each other. The question is whether the choice of the partition matters, over and above the fact that your total evidence will be drawn from a partition.

To take stock: I have considered two ways of justifying Relativisms: the hyperintensionalist account and the contrastivist account. Neither seems fully satisfactory. While I do not take my objections to foreclose the possibility that some version of Relativism may be true, I think it is also time to look for an alternative solution to the Borel-Kolmogorov paradox. I shall sketch one such solution in the remainder of this chapter.

4 Borel and Bertrand

Let’s start with a simple example.

Balls. Two balls A and B are drawn at random from the same urn with replacement. You know the urn contains 12 balls of three different sizes: 1, 10 and 10.1 (there is at least one ball of each size). Let s_A and s_B denote the sizes of A and B respectively. Let BIG be the event that $s_A \geq 10$, and SAME SIZE the event that $s_A = s_B$. How confident should you be in BIG, given SAME SIZE?

The answer is not obvious; it depends on your prior estimate of s_A and s_B , or equivalently the number of balls of each size. You might think, absent further information, your credence distribution of s_A and s_B should be uniform over the set $\{1, 10, 10.1\}$, in which case

$$C(\text{BIG}|\text{SAME SIZE}) = \frac{2}{3}.$$

Or you might think, given that 10.1 is very “close” to 10, they should be treated as approximately the same size. To this extent, you should assign equal credence to BIG and its negation. As a result,

$$C(\text{BIG}|\text{SAME SIZE}) = \frac{1}{3}.$$

Or you might think the answer should be a weighted average of the two, or something completely different. The answer does not matter for the purposes of this discussion. The crucial point is that this “paradox” is familiar – it is a version of the *Bertrand’s paradox*. When you encounter a pedestrian who is from Tokyo, Beijing or Shanghai, you may be 1/2 or 1/3 confident that they are from Japan rather than China, depending on how you group the possibilities. More generally, the Bertrand’s paradox illustrates the difficulty of addressing *the problem of priors*: In some epistemic situations, there does not *seem* to be one unique credence function that an agent must adopt, on pain of irrationality.

Note the emphasis on the seeming. I am not suggesting that there are multiple rational priors that you may adopt in *Ball*. Indeed, some theorists (call them “objectivists”) think that, despite appearances, there is always one unique rational credence function that an agent should adopt in a given epistemic situation – the credence function obtained by conditionalizing on the uniquely correct *ur-prior*. As a result, if you are 2/3 confident in BIG given SAME SIZE, whereas your twin with the exact same information is 1/2 confident in BIG given SAME SIZE, then one of you must be irrational. Perhaps an epistemologist would know who got it right by squinting at the problem hard enough; perhaps the answer is beyond our ken. Either way, the fact that it may not be obvious *to us* which one of you got it right is no conclusive reason to think that there is no right answer to begin with. We don’t know whether the Goldbach Conjecture is true; it does not necessarily follow that it is neither true nor false (or its truth is relative).¹⁶

I think, at its core, the Bore-Kolmogorov paradox is a version of the Bertrand’s paradox. We are torn about the value of $C(\text{FAST}|\text{SAME SPEED})$ for the same reason that we are torn about the value of $C(\text{BIG}|\text{SAME SIZE})$. In both cases, there are apparently competing considerations for what rational prior one should adopt, and the answer does not seem determinable by closer examination of the case. Crucially, *contra* Relativism, my diagnosis of the paradox does not entail that rational conditional credence given a null propositions *should* be undefined and/or relativized to the choice of a partition. As noted above, the fact that there does not *seem* to be one uniquely correct answer to $C(\text{FAST}|\text{SAME SPEED})$ does not entail that the question admits of multiple answers or no one at all. It shows that the uniquely correct answer is not obvious, if not unknowable. But the objectivist has conceded that charge long time ago. It is not easy to pinpoint “the magic probability function falling from the sky,” but we must embrace its existence on the basis of other theoretical considerations.

Now admittedly there is one obvious disanalogy between *Ball* and *Projectile*. In *Ball*, $C(\text{BIG}|\text{SAME SIZE})$ is determined by $C(.|\Omega)$ via the ratio formula. So the apparent indeterminacy of the former arises from the

¹⁶Some have conjectured that the Goldbach Conjecture may not be proven within ZFC, and its truth may to that extent be relative to our choice of axioms. However, that conjecture is not popularly received, and some remain hopeful that further reflection on the rationale of mathematical practice may help pin down one uniquely correct system of axiomatization (cf. Maddy (2017)).

apparent indeterminacy of the latter. On the other hand, in *Projectile*, $C(.|\Omega)$ is fully specified – it is the uniform Lebesgue measure over the unit square (where each point represents a possible combination of v_X and v_Y). Nevertheless, it appears that $C(\text{FAST}|\text{SAME SPEED})$ remains underdetermined.

Is this disanalogy relevant to the diagnosis of the case? I don't think so. Compare: Suppose in *Ball*, you know that there are exactly two balls of size 10. So the rational value of $C(s_A = 10|\Omega)$ is fully specified – it equals $1/6$. However, this information does not seem to settle what $C(.|\Omega)$ should be as a whole. Maybe you should split your credences between the two remaining possibilities; maybe you should be more confident that A is of size 1, in virtue of it being a more “distinct” possibility. Again, the answer does not matter. The point is: We don't seem to treat this variant as relevantly dissimilar from the original *Ball* case; both illustrate the difficulty of pinpointing the uniquely correct rational (unconditional) prior on the basis of the given. But if there is no relevant disanalogy between these two cases, then I contend that there is no relevant disanalogy between *Ball* and *Projectile*, either. In both cases, a partial specification of $C(.|.)$ does not determine the entire function as a whole.

Here is another way of stating the claim. There are two problems of the priors: (i) the problem of determining the unique rational *unconditional* credence function one should adopt in a given epistemic situation, and (ii) the problem of determining the unique rational *conditional* credence function one should adopt in a given epistemic situation. An answer to (ii) entails an answer to (i), but not necessarily *vice versa*. The Borel-Kolmogorov paradox reveals the residual difficulty of addressing (ii), even if we have a satisfactory answer to (i). While one could respond to that residual difficulty by giving up on the project of addressing (ii) altogether, it is unclear to me that someone who has already bitten the bullet about (i) should be particularly disturbed.

Before concluding, let me address one remaining issue. So far I have been assuming that rational agents have credal attitudes representable by a single probability function. But one might think it can be permissible if not required to have imprecise credences, i.e. credal attitudes representable by a set of probability function. Indeed, some objectivists take imprecision to be the solution to the problem of the (unconditional) priors. While there may not always be one unique rational precise unconditional credence function that one should adopt in a given epistemic situation, there is always one unique set of rational imprecise credence functions that one should adopt in a given epistemic situation. In *Ball*, maybe the uniquely rational set of probability functions include both $C(\text{BIG}) = 1/2$ and $C'(\text{BIG}) = 2$. As a result, $C(\text{BIG}|\text{SAME SIZE})$ should be $[1/3, 2/3]$, or some wider interval.

The philosophical status of imprecise credence is controversial and beyond the scope of this chapter. But setting those complications aside, I am skeptical that introducing imprecise credence would help with solving the problem of the *conditional* priors in general and the Borel-Kolmogorov paradox in particular. To see

why, recall the inconsistent triad of *Projectile*:

1. (Speed Continuity) $C(\text{FAST}|\text{SAME SPEED}) \approx C(\text{FAST}|\text{CLOSE SPEED})$
2. (Energy Continuity) $C(\text{FAST}|\text{SAME ENERGY}) \approx C(\text{FAST}|\text{CLOSE ENERGY})$
3. (Evidential Equivalence) $C(\text{FAST}|\text{SAME SPEED}) = C(\text{FAST}|\text{SAME ENERGY})$.

Both Speed Continuity and Energy Continuity seem compelling because, intuitively, eliminating a small margin of error (in either the difference in speeds, or the difference in energy) should not significantly change your credence in FAST. However, if that is true, then it would appear that eliminating a small margin of error should not significantly *dilate* your credence in FAST, either. That is, one might think $C(\text{FAST}|\text{SAME SPEED})$ should be precise or approximately precise, if $C(\text{FAST}|\text{CLOSE SPEED})$ is precise (especially if the respective ϵ is chosen to be extremely small). But since $C(\text{CLOSE SPEED})$ and $C(\text{CLOSE ENERGY})$ are both positive for any chosen $\epsilon > 0$, $C(\text{FAST}|\text{CLOSE SPEED})$ and $C(\text{FAST}|\text{CLOSE ENERGY})$ are both precise according to the ratio formula. It follows that $C(\text{FAST}|\text{SAME SPEED})$ and $C(\text{FAST}|\text{SAME ENERGY})$ must be two non-overlapping intervals; we are back to the same trilemma.

There is another, perhaps more pressing worry. Even if we set $C(\text{FAST}|\text{SAME SPEED}) = (a, b)$ for some $(a, b) \subset [0, 1]$, one might think that that interval should be “close” to the values of $C(\text{FAST}|\text{CLOSE SPEED})$ and $C(\text{FAST}|\text{CLOSE ENERGY})$ respectively (for the same reason that motivated Speed Continuity and Energy Continuity). Again, the notion of “closeness” between sets and numbers is vague and admits of different precisifications. But on an intuitive level, a set S can’t be “close” to a number t if *every* member of S is far away from t . So there should be some $x, y \in C(\text{FAST}|\text{SAME SPEED})$ such that $x \approx C(\text{FAST}|\text{CLOSE SPEED})$ and $y \approx C(\text{FAST}|\text{CLOSE ENERGY})$.

But there is nothing special about kinetic energy. The continuity principle seems just as compelling with respect to any quantity that is a strictly monotonic function of speed. By the result of Arnold and Roberston (2019) noted above, it follows that, for every $r \in (0, 1)$, there must be some $x \in C(\text{FAST}|\text{SAME SPEED})$ such that $x \approx r$. So $C(\text{FAST}|\text{SAME SPEED})$ must be the entire unit interval. This may seem repugnant to reason: If you already know that the speeds of the two projectiles differ by some small margin, how could learning that they are exactly equal to each other make your credence in FAST be *maximally imprecise*?¹⁷

To be clear, I do not take the above considerations to be decisive. After all, fans of imprecise credence must have already made peace with the idea that getting new evidence can sometimes lead to dilation rather than contraction. However, I think this shows that imprecision is at least not an obviously satisfactory

¹⁷Moreover, there is also nothing special about the event SAME SPEED. We could have run the same paradox with respect to the event that the speeds of the two projectiles differ by exactly am/s , for some $a \in (0, 1)$. It would appear that the imprecise theorist must think that, no matter what you learn as the difference between their speeds, your credence should dilate maximally. It follows that, not only your credence will dilate maximally, but that you can anticipate this with certainty.

solution to the problem of *conditional* priors. For if we understand conditional credence as the credence that one should adopt given the respective conditioning proposition (i.e. assuming Conditionalization), then what transpires from *Projectile* is that, in the dynamic setting, there can be competing considerations about *how imprecise one's credence should be*, in addition to what precise value it may assume.

5 Conclusion

I have argued that the Borel-Kolmogorov paradox is an instance of the more general problem – the problem of conditional priors. To this extent, the paradox no more supports the claim that rational conditional credence given null propositions should be undefined or relativistic, than any version of the Bertrand's paradox supports the conclusion that rational unconditional credence should be undefined or relativistic. Notably, this diagnosis suggests that the right response to the inconsistent triad of *Projectile* is to reject either Speed Continuity or Energy Continuity, rather than Evidential Equivalence. In *Ball*, it may be rational to be either $2/3$ or $1/2$ confident in BIG given SAME SIZE, but it would appear irrational to vacillate between these two conditional credences in the absence of new information. Similarly, in *Projectile*, it may be rational to be either $1/2$ or less than $1/2$ confident in FAST given SAME SPEED, but it does not follow that a single agent may potentially adopt both attitudes, depending on the specification of the subject matter of SAME SPEED or its contrasting alternatives.

Let me conclude by tying up a loose thread. Recall that, in Chapter 1 and Chapter 2, we considered two arguments that cause problems for CA: the question-begging objection (to the conglomerability argument for CA) and the argument against CA based on Weak Regularity. In each case, some CA theorists may get around the objection by rejecting the coherent theory of conditional probability in favor of Kolmogorov's theory, on the basis of the Borel-Kolmogorov paradox. What the paradox shows, they argue, is that the coherence norms for rational conditional credence are *weaker* than one might have thought. In particular, rationality does not demand Global Coherence, the principle that conditional credences given propositions drawn from different partitions should cohere with each other to form a connected whole.

However, the CA theorists' attitudes towards unconditional credences seem to be the exact opposite: The coherence norms for rational unconditional credences are stronger than one might have thought, i.e. they include CA. Moreover, rational unconditional credences should be coherent in this stronger sense, even if it means that the problem of unconditional priors becomes harder to solve. For a concrete illustration, consider a countable lottery. The anti-CA theorist might think that, given that ticket n is drawn from a countable lottery and nothing else, the rational probability that ticket n is the winning ticket is 0. The CA theorist cannot accept this answer. She must think that there is some finite set of tickets such that the probability

that the winning ticket belongs to this set is $1 - \epsilon$. What that set is may be an eternal mystery, but the CA theorist is willing to bite the bullet.

However, if I am right about the Borel-Kolmogorov paradox, i.e. it is an instance of the problem of (conditional) priors, then there is actually a tension between CA on the one hand, and the endorsement of Kolmogorov's theory of conditional probability on the basis of the Borel-Kolmogorov paradox on the other. For the CA theorist thinks that the problem of priors is the price one has to pay for coherence, whereas the Kolmogorovian thinks that coherence is the price one has to pay, in order to solve the problem of (conditional) priors.

Chapter 4

Conditioning and Crossing the Bridge

Every day is a new day.

Ernest Hemingway, *The Old Man and the Sea*

1 Introduction

The previous chapter left us with a puzzle. Recall the following two cases:

Three Prisoners. Percy, Bill and Roy are being tried for espionage. Two of them will be sentenced to life in prison; one will be set free. The person who will be pardoned is determined by a fair lottery, and only the prison guard knows who it is.

As they wait for the judgment day, Roy came up with an idea:

“Right now I am $1/3$ confident that I will be set free. However, I could ask the guard to name one person who will be condemned, other than myself. Whatever he tells me, it means that the person who will be pardoned is either me or the other person who is not named. So I’ve raised my odds from $1/3$ to $1/2$ – how smart!”

Opaque Three Prisoners. Same set-up as above, except that this time, the guard will communicate the name to Roy by telepathy instead of verbal communication. Moreover, a bad introspector that he is, Roy will acquire that information without realizing that he does. Again, Roy reasons to himself,

“Right now I am $1/3$ confident that I will be set free. However, I could ask the guard to name one person who will be condemned, other than myself. Whatever he tells me by telepathy, it

will mean that the person who will be pardoned is between me and the other person who is not named. So I've raised my odds from $1/3$ to $1/2$ – how smart!”

Intuitively, both Roys seem irrational. However, Conditionalization says that only one of them is – the Roy in *Three Prisoners*; the Roy in *Opaque Three Prisoners* (for simplicity, call him “Royo”) is rationally impeccable.

To dramatize the counter-intuitiveness of this verdict, consider a case where there are one million prisoners, only one (chosen at random) will be pardoned; the rest will be sentenced to life in prison. According to Conditionalization, Royo can rationally increase his confidence in the prospect that he will be freed from $1/1000,000$ to $1/2$, just by asking the guard to name 999,998 prisoners who will be imprisoned for the rest of their lives, other than himself.

On the other hand, it is not clear what Royo did wrong. We considered one possible explanation in Chapter 3: Royo is irrational because he fails to take into account what he could have learnt instead; rationality requires due sensitivity to actual as well as counterfactual evidence. Another tempting thought is that Royo is irrational because his updating plan is inconglomerable – he plans to change his credence in a predictable direction, no matter what he learns.

Neither explanation seems completely satisfactory. As discussed in Chapter 1, some form of rational inconglomerability seems ineliminable from our best theory of epistemology. Contrastivism also appears a last resort. Intuitively, we shouldn't change our mind, just by learning that we could have learnt something instead.

This chapter offers a different diagnosis. I argue that both Roy and Royo are irrational, though for different reasons. Roy is irrational because he fails to update on his total evidence. Royo is irrational because his updating plan (i.e. Bayesian conditionalization, or BCOND for short) is not *self-stable* – it recommends changing to a different plan, given information about itself. My primary goal is to introduce and motivate self-stability as a property of rational plans in general and rational updating plans in particular. My secondary goal is to relate self-stability to other important concepts in epistemology. In particular, I show that (i) an updating plan is self-stable if and only if it is immodest, i.e. it does not expect another updating plan to be more accurate than itself, and (ii) an updating plan is self-stable if its recommendation is worthy of deference. A corollary of my results is that, insofar as it is possible for BCOND to be non-self-stable (i.e. cases like *Opaque Three Prisoners* are metaphysically possible), there is a tension between Conditionalization on the one hand, and the rationality of immodesty/deference-worthiness on the other.

Plan. §2 introduce the notion of self-stability. §3-5 discuss how self-stability relates to other epistemological concepts. §6 conclude with some general philosophical reflections.

2 Self-Stability

2.1 Self-stable plans

We plan what to do on the basis of what we will learn. I plan to have cereal for breakfast if I find milk in my fridge, and toast otherwise. Formally, we represent a **learning experience** by a function E that maps a world w to a proposition $E(w)$, representing the total information the agent will learn at w from her learning experience. For example, suppose I will find milk in my fridge if and only if there is in fact milk in my fridge. Let M be the proposition that there is milk in my fridge. Then my learning experience of checking my fridge can be represented by the function:

$$E(w) = \begin{cases} M & \text{if } w \in M \\ \neg M & \text{if } w \notin M. \end{cases}$$

Fix a learning experience E . A **plan given E** corresponds to a function f_E that assigns each possible state $w \in W$ to an action $f_E(w)$, interpreted as the action to be performed at w given $E(w)$ (as well as any background information the agent might have).¹ For instance, my breakfast plan with respect to my fridge-checking learning experience E can be represented by the function

$$f_E(w) = \begin{cases} \text{have cereal,} & \text{if } w \in M \\ \text{have toast,} & \text{if } w \notin M. \end{cases}$$

In general, it seems rational to plan to do A if one will learn E , but plan to do something else, if one will learn E and something else in addition. For instance, suppose that, after I check my fridge, I will also check my cupboard to see if I have some overripe bananas left. If I do, then I may change my mind and make some banana muffins instead. Formally, let B denote the set of worlds where I have overripe bananas. Again, let's suppose that I will find overripe bananas if and only if there are overripe bananas in the fridge. In this example, my entire learning experience (which involves checking the fridge as well as the cupboard) is no

¹The definition of plan given here is slightly different from the one I gave in Chapter 2. There I represented a plan by a function from partitions to actions. It is straightforward to check that these two definitions are equivalent – a plan in the current sense induces a unique plan in the previous sense and *vice versa*. I adopt this alternative formalism for technical reasons that will become clear shortly.

longer representable by E , but the function E' given by

$$E'(w) = \begin{cases} B \wedge M & \text{if } w \in B \wedge M \\ \neg B \wedge M, & \text{if } w \notin B, w \in M \\ \neg B \wedge \neg M, & \text{if } w \notin B, w \notin M \\ B \wedge \neg M, & \text{if } w \in B, w \notin M \end{cases}$$

We can represent the plan relativized to this learning experience by the function $f_{E'}$:

$$f_{E'}(w) = \begin{cases} \text{have banana muffins,} & \text{if } w \in B \\ \text{have cereal,} & \text{if } w \notin B, w \in M \\ \text{have toast,} & \text{if } w \notin B, w \notin M \end{cases}$$

Given two learning experiences E and E' , let $E \wedge E'$ denote the function that maps each world w to $E(w) \cap E'(w)$. Intuitively, $(E \wedge E')(w)$ represents the total incremental information one gains at w after going through E and then E' (the order is assumed to not matter). My breakfast example suggests that, in general f_E may not be the same as $f_{E \wedge E'}$. Sometimes, it is rational to (plan to) adopt a new plan given new information.

However, sometimes it seems not. Consider the following example:

Grocery. Beth is in charge of grocery shopping today. Usually it's Ann's job, but she has to go pick up a friend at the airport. Knowing how clueless Beth is about e.g. the difference between red and green onions, Ann left Beth with a list of very detailed instructions, including things like:

Buy red onions if they have any. If they've run out of red onions but have some yellow ones left, buy those. If they've run out of both, then forget about onions.

Ann also put a sticky note on top of the list, which reads:

If you will learn whether you are in a situation where the list tells you not to buy any onions, then scratch the list, and buy some green onions if they've run out of red onions

If you were Beth, you may be baffled: What does Ann mean? Admittedly, her instructions are consistent; as noted above, it is fine to make different plans given new information. But her instructions also appear weird in a peculiar way. They tell Beth to do one thing, but something different once Beth learns what they tell her to do.

Grocery is a case about planning for other people. But the point generalizes to planning for oneself as

well. Suppose Beth is deliberating between various substitutions she will make if necessary before going to the grocery store, and she comes down to the following plan:

My first choice will be red onions. If they’ve run out of those, I’ll go with yellow onions. But if they’ve run out of both, I’ll do with no onions for today – except, if I’ll learn whether I am in a situation where I planned to not buy any onions, then I shall buy some green onions if they’ve run out of red onions.

Again, the plan is consistent. Formally, it is no different from my plan of having cereal if there is milk, toast if there is no milk, but banana muffin if I shall find out that I have overripe bananas. Yet Beth’s plan also appears defective in a way that my breakfast plan is not. It seems fine for me to plan to revise my plan, given information about what I have left in my cupboard; it seems irrational for Beth to plan to revise her plan, given information about what she has planned to do.

I think this reveals a general constraint on rational plans. Intuitively, rational plans should be *self-stable*, in the sense that they are not subject to revision given the truths about their own recommendations. If learning that you planned to A and nothing else should change your mind about A, then you should not have planned to do it in the first place. Plans appear defective, if they are effective when and only when one does not know what they tell them to do.²

Let us make the above intuition more precise. Fix a learning experience E and a plan f_E . Let E_{f_E} be the function that maps each world w to the proposition that f_E makes the recommendation $f_E(w)$, which corresponds to the set $[f_E = f_E(w)] = \{v : f_E(v) = f_E(w)\}$. Intuitively, $E_{f_E}(w)$ represents the total information one acquires at w after learning the true recommendation of f_E at w and nothing else. The above cases suggest the following necessary condition for rational plans:

Self-Stability. For any learning experience E , if f_E is rational for agent S , then S should not plan to revise it, given true information of its recommendation and nothing else, i.e. if f_E is rational for S only if

$$f_E = f_{E \wedge E_{f_E}}.$$

To get a feel for self-stability, let me contrast it with two other important notions in epistemology (I’ll spell out their connections more precisely in the next two sections).

First, say a plan (given E) is **modest** if it recommends a plan that makes different instructions than

²A quick caveat: Although I speak of self-stability as a constraint on rational plans, I think it is perhaps more natural to think of it as a norm for the activity of *rational planning*. The thought is, *when engaging in the activity of planning*, rational agents should not make two contingency plans, one as her default plan and the other as the plan that she would follow if she finds out what her first plan tells her to do.

itself; a plan is *immodest* if it is not modest.³ For instance, suppose Ann includes the following as one of her instructions:

In any case, follow the other list that I've put in your handbag.

Suppose in addition that this other list gives instructions inconsistent with the original list. Then Ann's instructions would be modest. Note that modest plans are inconsistent. If Ann includes the additional instruction, then Beth can't follow all of her instructions at the same time; one cannot plan to A and plan to follow a plan that says one shall not A.

As stated, modesty and non-self-stability differ in one crucial respect: unlike modest plans, non-self-stable plans are *prima facie* consistent. As noted above, it seems possible for Beth to plan to buy no onions if there are no red or yellow ones left, and also plan to buy green onions if she learns that her original plan tells her to buy no onions, just as it would be possible for me to plan to have cereals for breakfast if there is milk in the fridge, and to plan to have banana muffins instead if I find overripe bananas.

However, there is one caveat: in Beth's case, since Beth's original plan tells her to buy no onions if and only if there are no red or yellow ones left, it is possible for her plan to be non-self-stable only if it is possible for her to be ignorant of whether there are red or yellow onions left in the grocery store. While that may seem *prima facie* plausible – *a priori*, there is no reason why Beth should be omniscient about how many red or yellow onions are left at the grocery store – there may also be a sense in which that *shouldn't* be possible.

This brings us to the second property of plans that is closely related to self-stability. Recall that a plan can be thought of as a family of conditionals whose antecedents are mutually exclusive and jointly exhaustive (cf. Chapter 2). (Note that a plan that violates this constraint either reduces to a plan that satisfies it, or else gives conflicting recommendations at some states, and is defective to that extent.) Say a plan is **luminous** for an agent S if the antecedent of each of its conditionals is luminous for S at the time when the plan is to be executed; a condition C is luminous at time *t* for S just in case, if C obtains at *t*, then S knows (or be in a position to know) that C obtains at *t*. For instance, suppose I shall check the fridge thoroughly and my eyesights are reliable. Then my original breakfast plan would be luminous for me in the required sense: there is milk in the fridge if and only if I will know that there is.

Some think all rational plans should be luminous (Schafer 2018). Call this view **Luminosity**. The thought seems to be that rational agents should not plan to act differently under circumstances that are indistinguishable from their own subjective perspectives. Suppose I am about to make a half-boiled egg. Intuitively, it would not be a good idea to plan to turn off the heat, if and only if the yolk starts to set.

³Lewis (1971) introduced *immodesty* as a property of inductive methods. A method is *immodest*, according to Lewis, just in case it recommends itself, in the sense that it expected itself to be most accurate. Here I am adapting that definition to the case of plan. Indeed, as I show in the next subsection, there is a precise sense in which self-stable updating plans are immodest in the Lewisian sense.

Objectively speaking, the plan is impeccable. But given that I can't see through egg shells, it makes a terrible plan from the point of view of execution.

One might think Luminosity explains the intuitive irrationality of Ann's plan for Beth in *Grocery*. As noted above, Ann's instructions are consistent only if it is possible for Beth to not know whether there are red or yellow onions left when she goes to the grocery store. But that means her original plan for Beth is non-luminous. It tells Beth to buy no onions if there are no red or yellow ones left, and yet Beth may not know whether that condition obtains.

Luminosity implies Self-Stability, assuming that the agent knows what the plan is (as a family of conditionals from circumstances to actions) and her knowledge is closed under logical entailment. To see how: take a plan f . Suppose f is luminous, i.e. whenever a relevant circumstance C obtains, the agent will know that C obtains (when it is time for her to execute the plan). Since the agent knows that f is the plan and her knowledge is closed under entailment, she will know that C obtains only if she will know that f tells her to $f(w)$, for some $w \in C$ (the choice of w does not matter, as f is supposed to give the same recommendation across states where C obtains). Therefore, when C obtains, the agent will know what she should do in that circumstance, according to the plan. Since this applies to all possible circumstances that feature as antecedents of f , it follows that, whenever f tells the agent to A , the agent will know that f tells her to A . Presumably, learning what one already knows should not make one revise one's plan. So f is self-stable. (I'll give a more formal proof in the next section.)

On the other hand, Self-Stability does not imply Luminosity. One may not plan to revise one's plan, even if one could be ignorant of whether/when some of the relevant circumstances obtain. I plan to have snacks between meals if and only if I am hungry. The plan is non-luminous, for I do not always know whether I am hungry. Nevertheless, I may plan to stick to the plan, if I learn what it tells me to do. In fact I may welcome that piece of information – it would help me follow my original plan.

So Luminosity is a stronger condition than Self-Stability. The question is: do rational plans have to be luminous, over and above being self-stable?

I am skeptical for two reasons. First, I share the externalist worry that there may be no necessarily luminous conditions (cf. Williamson (2000) and Srinivasan (2020)). Back to *Grocery* again. Clearly the condition that *the store has run out of red and yellow onions* is not necessarily luminous. But one might think Ann's original instructions are sloppy. What she should have told Beth is that, if *she learns that* there are no red or yellow onions left, then buy no onions. However, it is far from clear that learning is luminous in the required sense either. For instance, one might think learning that p is a matter of being convinced that p is true. If that is right, then learning is luminous only if we are always certain of our own certainties. But both empirical studies and introspection (not unironically – as Srinivasan (2020) points out) suggest

that our epistemic access to our own mental states is not necessarily better than our epistemic access to the number of onions left in a grocery store. We are not necessarily better at looking inwards than outwards.

My second concern is more flat-footed. I think the motivation for imposing luminosity is followability: a plan is rational only if it is followable, and a plan is followable, the thought goes, only if it is luminous, i.e. it does not give different recommendations under different circumstances that the agent cannot tell apart based on what she knows. But one might think, in order to follow a plan that says *if C then A*, one only needs to be *sensitive* to whether C obtains, in the sense that one can act differently in response to the truth of C. If I go to bed at a proper time, I can follow the plan of waking up between 7-8am in the morning, without setting an alarm clock. I may not know that it is between 7-8am when I wake up, and I certainly don't before I wake up. Yet there seems to be nothing defective about my wake-up plan. If this is right, then rational plans do not have to be luminous, even if they have to be followable; followability presupposes sensitivity rather than knowledge. (Note that this suggests an alternative diagnosis of the defectiveness of my plan for making a half-boiled egg. It would be irrational to plan to turn off the heat if and only if the yolk starts to set, not because I would not *know* whether the yolk is beginning to set, but because, as a matter of biological fact, I am not *sensitive* to what is happening inside the egg shell.)

Before concluding this section, let me introduce one more technical notion: say a plan f is **at least as available as** f' just in case it does not give different recommendations at two states where f' tells the agent to do the same thing, i.e. $f'(w) = f'(v) \Rightarrow f(w) = f(v)$. Intuitively, if we understand followability in terms of sensitivity, then f is at least as available as f' means that an agent can follow f if she can follow f' .

2.2 Self-stable updating plans

So far we have been talking about plans in general. Let us turn now to updating plans more specifically. Informally, an **updating plan** is a plan of a special kind. Instead of specifying what to do, it specifies what credence function to adopt at a future time, after a course of learning experience. More precisely: fix a set of states W . Let C be an unconditional credence function over W , representing the agent's prior at t_0 . For simplicity, I'll also assume that the agent's prior C is always regular and well-defined over some sigma-algebra $\mathcal{F}_C$ (recall that regularity means $C(w) > 0$ for all $w \in W$. In particular, conditional credences are always well-defined by the ratio formula.) Let $\Delta(W)$ denote the set of unconditional, probabilistic credence functions over W . An updating plan given a learning experience E is a function of the form

$$f_E : W \rightarrow \Delta(W)$$

where $f_E(w)$ represents the credence function that the agent plans to adopt at w at t_1 , after learning information $E(w)$ and nothing else.

Recall that self-stability says: f_E is rational (for a subject S) only if $f_E = f_{E \wedge E_{f_E}}$. Before I did not say what $f_{E \wedge E_{f_E}}$ is. In general that would depend on the agent's preferences given the true partition element of $\mathcal{E}$. When I discover that there are bananas left in the fridge, I may stick to my original plan if I don't like bananas, or change my mind if I dislike wasting food more.

However, in the case of updating, there is a natural candidate for what $f_{E \wedge E_{f_E}}$ should be: for each w , $f_{E \wedge E_{f_E}}(w)$ should equal the credence function recommended by f_E at w , namely $f_E(w)$, conditional on $f_E = f_E(w)$, i.e.

$$f_{E \wedge E_{f_E}}(w) = f_E(w)(\cdot | f_E = f_E(w)).$$

In particular, an updating plan f_E is **self-stable** just in case it is stable under conditionalization on its own recommendations, i.e.

$$f_E(w)(\cdot) = f_E(w)(\cdot | f_E = f_E(w)), \forall w \in W.$$

While I have some reservations about this way of defining self-stability with respect to updating plans – in particular it assumes conditionalization – I shall assume it for the rest of this chapter. (I shall also drop the subscript E when it is clear from the context.) This allows us to restate the principle of Self-Stability for updating plans more succinctly as follows:

Self-Stability. An updating plan f is rational only if $f(w)(\cdot) = f(w)(\cdot | f = f(w)), \forall w \in W$.

Before proceeding, let us work through the case of *Opaque Three Prisoners*. Recall that Royo's evidential situation is given by:

$$E(R \& T_P) = E(B) = \{R \& T_P, R \& T_B, B\}, E(R \& T_B) = E(P) = \{R \& T_P, R \& T_B, P\}.$$

Suppose Royo starts out with a uniform prior. Let f be the BCOND-plan given his evidential situation. As noted before, no matter what the guard tells him, f would recommend Royo to be 1/2 confident that he will be freed. However, this plan is not self-stable. For instance, conditional on the information *that his plan says he should be 1/2 confident that Bill will be freed*, Royo should plan to be less than 1/2 confident that he will be freed. This is because only in some cases where he will be freed would his plan recommend him to be 1/2 confident that Bill will be freed; in some others, his plan will tell him to be 1/2 confident that Percy will be freed. As a result, learning what his plan tells him to do rules out some of the worlds where he will be freed, but none of the worlds where Bill will be freed, which means it is evidence against his own

prospect of freedom. Therefore, if Royo plans to update by BCOND, then *ex ante*, he is certain that he will deviate from the plan, once he learns what the plan tells him to do.

In the next three sections, I shall spell out in more detail how Self-Stability relates to the issue of luminosity, immodesty and epistemic deference. My primary goal is not to defend the principle, but to clarify what fruits it may bear, and what would be at stake, if a debate is to be had.

3 Luminosity

3.1 Plan Luminosity

Recall that a plan is luminous for an agent S at time t just in case the antecedents of the conditionals that make up the plan are each luminous for S at time t , i.e. if the plan says “if C then A ,” then S knows/is in a position to know at t that C obtains, if it obtains.

For present purposes, say S knows p just in case p is entailed by S ’s total evidence (so evidence is factive and propositional). Then we can define plan luminosity more rigorously as follows: A plan f_E is **luminous** just in case the truth of its recommendation is entailed by the agent’s incremental evidence after undergoing the learning experience E at w , i.e.

$$E(w) \subseteq [f_E = f_E(w)], \forall w \in W.$$

What we observed in the previous section is that self-stability follows from luminosity plus evidential certainty. (Recall that an updating plan f_E is evidentially certain if, for every w , $f_E(w)(E(w)) = 1$.) To spell out the argument in more detail: Suppose f_E is luminous and evidentially certain. Then for every w , $f_E(w)(f_E = f_E(w)) \geq f_E(w)(E(w)) = 1$. So $f_E(w)(\cdot) = f_E(w)(\cdot | f_E = f_E(w))$, i.e. f_E is self-stable. Note that the assumption of evidential certainty corresponds to the assumption implicit in the last step of my previous argument, namely that updating on information that one already knows should not lead to rational revision of one’s plan.

As mentioned above, luminosity is strictly stronger than self-stability. Before we gave an informal example; now we can be more precise. Consider the omniscient plan

$$f_E(w) = v_w, \text{ where } v_w(A) = \begin{cases} 1 & \text{if } w \in A \\ 0 & \text{if } w \notin A. \end{cases}$$

The omniscient plan is always self-stable, since $[f_E = f_E(w)] = \{w\}$ and $v_w(\{w\}) = 1$. On the other hand,

this plan is luminous when and only when the agent is in fact omniscient, i.e. her evidence at w corresponds to the singleton $\{w\}$.

One last remark before ending this section: the crucial step in my proof of the entailment

$$\text{E-luminosity} + \text{E-evidential certainty} \Rightarrow \text{self-stability}$$

is that the two conditions jointly imply the following property of updating plans:

$$\forall w \in W, f(w)(f = f(w)) = 1.$$

Huttegger and Nielsen (2020) named this property “luminosity.” To avoid terminological confusion, I shall call it “self-luminosity.” Informally, an updating plan f is **self-luminous** just in case it never recommends assigning positive credence to the event that it recommends updating differently than it does. We have seen that self-luminosity implies self-stability. It is straightforward to check that the other direction holds as well: if $f(w) = f(\cdot | f = f(w))$, then $f(w)(f = f(w)) = f(f = f(w) | f = f(w)) = 1$. So an updating plan is self-stable if and only if it is self-luminous. Equivalently, Self-Stability as a rational norm is equivalent to the following principle:

Self-Luminosity. An updating plan f is rational only if $f(w)(f = f(w)) = 1, \forall w \in W$.

On the other hand, self-stability is also more general than self-luminosity. For a plan f can be self-luminous only if $f(w)$ is a credence function for every $w \in W$, i.e. it is an updating plan. On the other hand, self-stability is a property of plans in general. My plan to have snacks between meals if and only if I am hungry is self-stable (whereas it would not make sense to call this plan self-luminous). Since I think updating plans should be self-stable to the extent that all plans should be self-stable, I take Self-Stability to be normatively more fundamental (in an intuitive, non-technical sense) than Self-Luminosity.

3.2 Evidential Luminosity

We can extend the notion of luminosity from plans to learning experiences. Intuitively, a learning experience is luminous if, no matter what the agent will learn, it is part of her evidence what her evidence is. More precisely, **E** is **luminous** if

$$\text{E}(w) \subseteq [\text{E} = \text{E}(w)], \forall w \in W.$$

where $[\text{E} = E] = \{v \in W : \text{E}(v) = E\}$ represents the proposition *The agent’s total evidence is E*. Notably, assuming that evidence is factive, i.e. $w \in \text{E}(w)$, a learning experience is luminous if and only if $\text{E}(w) = [\text{E} = \text{E}(w)]$ for all w , which is in turn equivalent to the set $\{\text{E}(w)\}_{w \in W}$ being a partition of W .

Fix a learning experience E . Let $\mathcal{E}$ denote the partition $\{[E = E(w)]\}_{w \in W}$. We say an updating plan f_E is **E-available** if it does not tell the agent to adopt different credence functions at two states where her total evidence is the same, i.e. $E(v) = E(w) \Rightarrow f(v) = f(w)$. We say f_E is **evidentialist** if at every w , $E(w)$ is the strongest proposition that is assigned credence 1 by $f_E(w)$.

Fact 1. *If f_E is E-available and E-evidentialist, then f_E is self-stable if and only if E is luminous.*⁴

As a special application of Fact 1, note that the BCOND-plan $f_E : w \mapsto C(\cdot | E(w))$ is E-available and E-evidentialist whenever C is regular. So the BCOND-plan is self-stable when and only when the relevant learning experience is luminous (assuming that the agent starts out with a regular prior). In particular, the BCOND-plan is not self-stable in the case of Royo, whose evidential situation is non-partitional and thereby non-luminous.

It follows from Fact 1 that, assuming (i) updating plans should be available to their respective evidential situations – rational agents should not plan to update differently in response to the same piece of evidence and (ii) one should only plan to assign credence one to one’s total evidence, Self-Stability is equivalent to the following metaphysical thesis about evidence:

Evidential Luminosity. Necessarily, evidence is luminous.

Whether this is good or bad news for Self-Stability depends on one’s view about Evidential Luminosity. We have seen one putative counter-example to the principle – *Opaque Three Prisoners*. Here are two more mundane cases:

Red Wall. You are visiting the Forbidden City for the first time. According to the map, if you turn right at the next corner, you will hit a wall that *looks* red. Suppose you know that the map is perfectly reliable. Based on what you know about the Forbidden City, you are 0.5 confident that the wall will be red and look red to you, and 0.5 confident that the wall will be white but look red to you (perhaps because it will be lit up with red light).

Unmarked Clock. You are trying to catch a bus that is supposed to arrive at 9:50am. You arrived at the bus stop, but there is no sign of the bus. Unfortunately your phone has run out of battery, and you’ve left your watch at home. So all you can do is squinting at a clock on a shopping mall nearby. To make matters worse, the clock is extremely austere – it only has a clock hand with no markings.

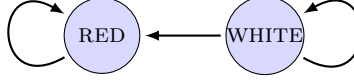
Some people think that, in *Red Wall*, you get evidence that the wall is red if it is in fact red, but you do not acquire any new information if it is in fact white, i.e. your evidential situation is representable by the

⁴I include all the proofs of Facts in Appendix C.

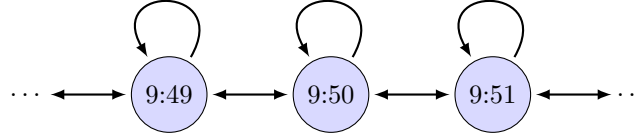
following function:

$$E(R) = \{R\}, \quad E(W) = \{R, W\}$$

which is non-luminous. In particular, $E(W) \not\subseteq [E = E(W)] = \{W\}$.



Red Wall is a case where, possibly, you might lack evidence of what your evidence is. *Unmarked Clock* seems to demonstrate a more radical failure of Evidential Luminosity: in that case, necessarily, you lack evidence of what your evidence is. For presumably, no matter how accurate your vision is, you can't tell precisely which of the 720 possible positions at which the clock hand is pointing. Perhaps the best you can do is to make an estimate with at least one minute margin of error. For instance, if it is 9:50am, then you can tell that it is past 9:48am and before 9:52am, but you can't get more precise than that. But then your evidence is necessarily non-luminous. For every t , $E(t) = \{t - 1, t, t + 1\} \not\subseteq [E = E(t)] = \{t\}$.



Whether Evidential Luminosity is true remains an issue of controversy. Some take cases like *Opaque Three Prisoners*, *Red Wall* and *Unmarked Clock* to be decisive counter-examples to Evidential Luminosity (Williamson 2000; Lasonen-Aarnio 2015). Others think the cases are either metaphysically impossible or misdescribed (Stalnaker 2015; Salow 2018; Berker 2008). Perhaps we should have chosen a richer sample space. Perhaps the model is true only of agents who are irrational in some non-obvious sense.

For what it's worth, I am inclined to reject Evidential Luminosity. For one thing, the principle is intimately connected to Luminosity (all rational updating plans should be luminous), a principle of which I am skeptical. To spell out their connection more precisely, recall that Evidential Luminosity says

$$E(w) \subseteq [E = E(w)], \forall w \in W$$

whereas Luminosity says

$$E(w) \subseteq [f = f(w)], \forall w \in W.$$

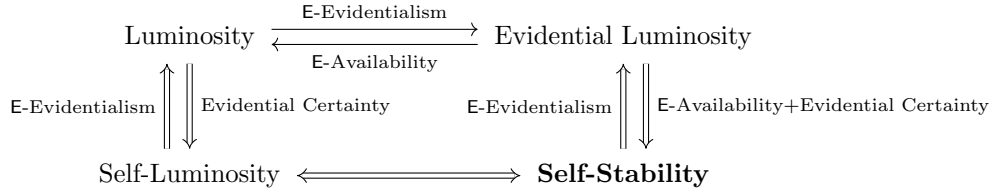
So Luminosity entails Evidential Luminosity whenever the relevant updating plan is evidentialist with respect to E (since E -evidentialism entails that $[f = f(w)] \subseteq [E = E(w)]$), and Evidential Luminosity entails

Luminosity whenever the relevant updating plan is E-available (i.e. $[E = E(w)] \subseteq [f = f(w)]$).

But more importantly, I think the reasons for rejecting Luminosity also speak against Evidential Luminosity. If we think of evidence as a piece of information that the agent has (in some sense), then *a priori* there seems to be little reason to suppose that evidence must be self-revealing in the way prescribed by Evidential Luminosity. There are many things that we have but lack evidence that we have them; it is unclear why evidence should be somehow special. On the other hand, if we think of evidence in subjective terms, e.g. the strongest proposition to which one assigns absolute certainty – then Evidential Luminosity amounts to the thesis that one must be certain of one’s own certainties. Again (not nonironically) introspection suggests that we are not necessarily good at introspecting.

Admittedly, this leaves me in an awkward position; I reject Evidential Luminosity but endorse Self-Stability, and yet Fact 1 shows that the two are in some sense equivalent. My response is to reject one of the assumptions needed for that equivalence to hold: E-evidentialism. I think it can be rational to plan to assign credence one to propositions that are strictly stronger than one’s total evidence. Where that leaves us with respect to the principle of total evidence is a question that I’ll take up in §6.

Here is a diagram that summarizes the results of this section:



4 Immodesty

Recall that a plan is modest if it recommends a different plan. In §2 we contemplated one possible definition of recommendation: f recommends f' just in case f says one should follow the recommendation of f' . More precisely, for all w , $f(w) = f'(w)$. This is why modest plans are inconsistent; f cannot equal both $f(w)$ and $f'(w)$ at w , assuming that the two actions cannot be performed at the same time.

However, when Lewis introduced the notion of immodesty with respect to inductive methods, he had a slightly different, weaker definition of “recommendation” in mind:

[A]n inductive method C recommends an inductive method C' if the C -mean estimate of the accuracy of C' is not exceeded by the C -mean estimate of the accuracy of any rival method.

While there seems to be no straightforward analogue of this definition for general plans – actions cannot be more or less accurate, or estimate the accuracy of other actions – the definition does appear to generalize to updating plans. Following Lewis, we may say an updating plan f **recommends** f' just in case f expects

f' to be the most accurate updating plan. To make this notion more precise, however, we need to define expected accuracy of one updating plan with respect to another, to which I shall now turn.

Let's start with expected accuracy of credence functions. Formally, an **accuracy scoring rule** is a function $\mathcal{A}$ that assigns real numbers to pairs of worlds and credence functions. Intuitively, $\mathcal{A}(C, w)$ represents how accurate C is at w – how close it is to the omniscient credence function v_w . The expected accuracy of one credence function Q with respect to another credence function P is then calculated by

$$\mathbb{E}_P[\mathcal{A}(Q)] = \sum_w P(w) \mathcal{A}(Q, w).$$

We say P **recommends** Q (**relative to** $\mathcal{A}$) if it expects Q to have the higher accuracy score according to $\mathcal{A}$ than itself does; P is **immodest** (**relative to** $\mathcal{A}$) if it expects itself to have the highest accuracy score according to $\mathcal{A}$.

Now in some sense, every credence function P is a trivial updating plan – the plan that says *Adopt P no matter what comes*. This perspective suggests a natural generalization. Let f and f' be two updating plans. We can define **the expected accuracy score of f' with respect to f** as the following quantity:

$$\mathbb{E}_f[\mathcal{A}(f')] = \sum_w f(w)(w) \mathcal{A}(f'(w), w).$$

Recall that $f(w)$ is the credence function that one will adopt at w , if one updates in accord with f . So $f(w)(w)$ represents the credence that one will assign to w , if one updates in accord with f . Thus, more informally, the quantity $\mathbb{E}_f[\mathcal{A}(f')]$ represents the weighted average of the accuracy scores of the credence functions recommended by f' – weighted by one's future credence in each state, if one updates in accord with f .⁵ It is also straightforward to verify that this notion of expected accuracy reduces to the standard definition if one views credence functions as constant updating plans.⁶ We say f **recommends** f' just in case

$$\mathbb{E}_f[\mathcal{A}(f')] > \mathbb{E}_f[\mathcal{A}(f)].$$

What would immodesty mean, on this alternative notion of recommendation? One naive idea is to say that f is immodest if it recommends no updating plans other than itself. But a moment of reflection suggests that this would not work, since every plan must recommend the omniscient plan. A better alternative is to

⁵Note that in general the diagonal function $w \mapsto f(w)(w)$ would not be a probability function. For instance, the diagonal function associated with the omniscient plan is the constant function that maps every state to 1.

⁶For completeness: let f, f' be two constant updating plans with $f(w) = P$ and $f'(w) = Q$. Then

$$\mathbb{E}_f[\mathcal{A}(f')] = \sum_w f(w)(w) \mathcal{A}(f'(w), w) = \sum_w P(w) \mathcal{A}(Q, w) = \mathbb{E}_P[\mathcal{A}(Q)].$$

say that f is **immodest** just in case it does not recommend any plans that is at least as available as f itself, i.e. for every f' such that $f(w) = f(v) \Rightarrow f'(w) = f'(v)$ for all $w, v \in W$,

$$\mathbb{E}_f[\mathcal{A}(f)] \geq \mathbb{E}_f[\mathcal{A}(f')]$$

with equality if and only if $f = f'$. The thought is, a rational updating plan should not recommend switching to another plan, provided that the agent can make the switch. It is rational to stick to the suboptimal, when and only when the optimal is beyond one's reach.

So we have a precise definition of when updating plans are immodest. The notion is empty, however, unless we say more about what we mean by accuracy scoring rules. Clearly, not any function that assigns real numbers to credence-world pairs makes an accuracy scoring rule; the function should also satisfy some constraints that accord with our pre-theoretical intuitions. One such constraint that many find attractive is *strict propriety*:

Strict-Propriety. An accuracy scoring rule does not make rational credence functions modest. In particular, it does not make any probabilistic credence function immodest, i.e. if $\mathcal{A}$ is an accuracy scoring rule, then for any probability function P and credence function Q ,

$$\mathbb{E}_P[\mathcal{A}(P)] \geq \mathbb{E}_P[\mathcal{A}(Q)]$$

with equality if and only if $P = Q$.

Most prominent scoring rules (e.g. the Brier score) are strictly proper. It turns out that, assuming Strict Propriety, self-stability is equivalent to immodesty for updating plans in the sense defined above.

Fact 2. *An updating plan f is self-stable if and only if it is immodest.*

Many think that rational credence functions should be immodest. Just as rationality requires one to believe the truth when they know what it is, rationality also requires one to adopt the most accurate credence function, when they know what it is. *Prima facie*, the same consideration generalizes to updating plans as well. Insofar as it is irrational to adopt self-undermining credence functions, it seems irrational to adopt self-undermining updating plans.

So Fact 2 suggests another argument for Self-Stability:

P1 Immodesty is a rational requirement for credence functions.

P2 If immodesty is a rational requirement for credence functions, then immodesty is a rational requirement for updating plans.

P3 An updating plan is immodest if and only if it is self-stable.

C Self-stability is a rational requirement for updating plans.

To be clear, I do not endorse this argument myself. For one thing, I prefer to think of self-stability as a rational requirement on *all* plans, not just updating plans. So a successful argument for Self-Stability by my lights should be similarly general in scope.

For another, the equivalence between self-stability and immodesty holds only under the assumption of Strict Propriety. Strict Propriety is often motivated in two steps: (i) rational credence functions should be immodest; (ii) every probabilistic credence function can be rational given the right circumstances (cf. Joyce 2009).⁷ However, as Blackwell and Drucker (2019) point out, both steps can be resisted. In particular, one who thinks there is more to rationality than mere probabilistic coherence may also think that some probabilistic credence functions are necessarily irrational. For instance, one may think under no circumstances is it rational to be confident that a coin will land heads, given that it has bias $1/2$. If this is right, then the true accuracy score only needs to make some but not all probabilistic credence functions immodest, which would in turn block the proof of Fact 2.⁸

5 Reflection

Lastly, Self-Stability also bears an intimate connection to a principle that we have encountered before – Reflection (cf. Chapter 2). Recall that the principle says, roughly, one should be disposed to adopt one’s future credence if one knows what it is. While this principle is sometimes understood as a synchronic constraint on rational priors, as many have pointed out (Easwaran (2013b), Huttegger (2013), and Pettigrew (2020)), the principle is best interpreted as a wide-scope constraint on *combinations of prior and updating plans* in the following sense:

Reflection. For any prior C and updating plan f , (C, f) is rational only if

$$C(A|f(A) = x) = x, \forall A \in \mathcal{F}_C, \forall x \in \mathbb{R}.$$

⁷More precisely: fix a sample space W . Let $\mathcal{C}(W)$ be the set of possible chance distributions over W . The thought is: $\mathcal{C}(W) = \Delta(W)$ – every probability measure could potentially be a chance measure and *vice versa*. Let C' be a rational credence function over $W \times \mathcal{C}(W)$. Assuming the Principal Principle (i.e. chance is the guide for rational credence), $C'(\cdot|ch = C)$ is rational and equals C when restricted to W .

⁸It may be worth noting that, if we restrict our attention to rational updating plans, then we do not need the full strength of Strict Propriety for proving the entailment from self-stability to immodesty. Instead, the following weaker assumption will do:

Weak Immodesty. A rational updating plan only recommends immodest credence functions.

That is, under Weak Immodesty, self-stability implies immodesty *under no assumptions about the accuracy scoring rule*. Moreover, the converse also holds if we assume in addition that f^* is always rational whenever f is.

In particular, if one violates Reflection, then *either* one's prior is irrational, *or* one's updating plan is defective. This formulation then evades some putative counter-examples to the principle, understood as a constraint on rational priors alone. For instance, suppose I had a croissant this morning (cf. Fraassen (1995)). It seems rational for me to be certain now that I had a croissant today. It also seems rational for me to be certain now that a week later I will be only 0.8 confident that I had a croissant today. However, it would appear irrational for me to be certain that I had a croissant today while *planning* to be 0.8 confident that I had a croissant today a week later. It is not wrong to forget, but it seems wrong to plan to forget (at least from an epistemic point of view).

For the purposes of this discussion, however, I shall focus on a stronger Reflection principle that I call *Global Reflection*. Fix a sample space W , a credence function C and an updating plan f . We say C **defers to** f just in case, for any credence function P , if f says that C should be revised to P , then C is equal to P conditional on that information and nothing else, i.e.

$$f(w) = C(.|f = f(w)), \forall w \in W.$$

Global Reflection. Rational agents should defer to their updating plans, i.e. for any prior C and updating plan f , (C, f) is rational only if C defers to f .

Reflection says that, for any proposition, one should be disposed to adopt any credence that one plans to assign to that proposition; Global Reflection says that, for any credence function, one should be disposed to adopt that credence function if it is recommended by one's updating plan. Clearly Global Reflection entails Reflection. Moreover, it is strictly stronger. Here is a simple model that shows Reflection does not imply Global Reflection: Let $W = \{a, b, c\}$. Suppose C is the uniform prior, i.e. $C = (1/3, 1/3, 1/3)$. Let f be defined as follows:

$$\begin{cases} f(a) = (1/2, 1/2, 0) \\ f(b) = (0, 1/2, 1/2) \\ f(c) = (1/2, 0, 1/2). \end{cases}$$

One can check that $C(A|f(A) = x) = x$ for all $A \subseteq W$. On the other hand, $C(\{a\}|f = f(a)) = C(\{a\}|\{a\}) = 1 \neq f(a)(\{a\}) = 1/2$. So $C(.|f = f(a)) \neq f(a)$ (note that f is not self-stable either in this example).

Fact 3. *If C defers to f , then f is self-stable.*⁹

So Global Reflection entails Self-Stability, though the converse is not true. For instance, suppose one starts out with the prior C . Consider the plan that says *Switch to $C'(\neq C)$ no matter what comes*. This plan is

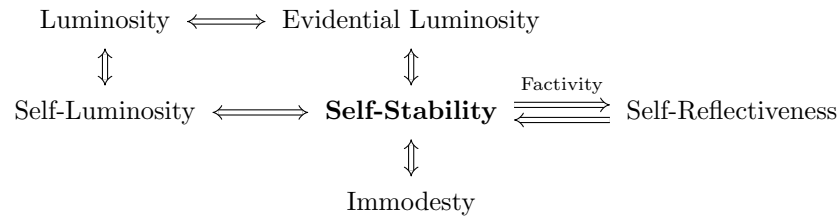
⁹The converse is true if (C, f) satisfies the principle of New Reflection (Elga 2013): $f(w)(.|f = f(w)) = C(.|f = f(w))$.

self-stable, but clearly C does not defer to f in the above sense.

Moreover, it is possible for BCOND to be self-stable but fail to be self-reflective. In fact, this happens, if and only if the relevant evidential situation is non-factive.

Fact 4. *If f is E-available, E-evidentialist and self-stable, then f is self-reflective only if E is factive. The converse is true if f is the BCOND-plan.*

The following diagram summarizes the logical relationship amongst various notions in the special case where f is the BCOND-plan:



Since the BCOND-plan is not self-stable whenever the agent's evidence is non-luminous, it follows that Global Reflection is in conflict with Conditionalization as well, if Evidential Luminosity is false.

Some take this as an argument against Global Reflection. Williamson puts the objection eloquently,

From our present perspective, the non-partitional structure of our future knowledge is a source of bias, similar in effect to forgetting although much subtler in its operation. Of course, we shall probably know more tomorrow, and it would be foolish then to disregard the new knowledge. But we cannot take advantage of the new knowledge in advance. We must cross that bridge when we come to it, and accept the consequences of our unfortunate epistemic situation with what composure we can find. Life is hard. (p.237)

It's worth noting that Williamson was targeting Global Reflection (or variants thereof) as a constraint on rational priors. As noted above, I prefer to think of such principles as constraints on rational combinations of priors and updating plans. I shall say more about the relationship between Global Reflection and Conditionalization in the next section as well as Chapter 5. At this point, it suffices to note that, while *cross that bridge when one gets there* may be alright from the perspective of deliberating what to do/believe here and now, it does not seem rationally recommendable from the perspective of planning. As a planning agent, it is your duty to deliberate carefully well in advance which bridge to cross, and how to cross it. (Spoiler: I think the planning agent faces a slightly different predicament – she must plan to take advantage of not only her new knowledge, but also information beyond what she will know. It is an ambitious plan, but she must give it a try.)

To take stock: I have shown that (i) f is self-stable if and only if it is immodest (in a precise sense), and (ii) C defers to f (or equivalently f is self-reflective) only if f is self-stable. Both results lend support to my original diagnosis of *Opaque Three Prisoners*: Royo is irrational because his updating plan is not self-stable; he has one plan about how he will update given the guard’s message, and a different plan, once he learns what is the credence function recommended by his original plan.

6 Crossing the bridge

But the hard-core Bayesian might remain unpersuaded. I motivated Self-Stability on intuitive grounds. But insofar as the principle seems intuitive, so does Conditionalization. While it is bad news that the two cannot both be true if cases like *Opaque Three Prisoners* are possible, one may take this as evidence against Self-Stability rather than Conditionalization. Unlike his counterpart, Royo is doing the best that he can as an epistemic agent. We must embrace the normative asymmetry between the two with composure, just as we have to accept that some line must be drawn amongst Debby, Lotti and Corinne (cf. Chapter 1).

For a start, I think Conditionalization is less intuitive than many have thought. The principle is often motivated by appealing to instrumentalist considerations, e.g. the BCOND-plan is the unique updating plan that avoids Dutch books and maximizes expected accuracy. But as we shall see in the next chapter, these arguments do not go through in cases like *Opaque Three Prisoners*, where the BCOND-plan is Dutch-bookable and does not maximize expected accuracy. So, insofar as these considerations go, they speak *against* Conditionalization as a universal rational norm, rather than in its favor.

Moreover, Conditionalization (as a norm for rational updating plans) is equivalent to the conjunction of the following two constraints:

1. Evidential Certainty: One should plan to be confident in one’s total evidence, i.e.

$$\forall w, f(w)(E(w)) = 1;$$

2. Rigidity: One should plan to retain one’s conditional credence given one’s total evidence, i.e.

$$\forall w, f(w)(\cdot|E(w)) = C(\cdot|E(w)).$$

While Evidential Certainty seems intuitive (cf. Chapter 2), it is not immediately clear why rational updating plans should be rigid. In fact, it might appear that in some learning situations, rational agents should update in accord with *non-rigid* plans, e.g. when accommodating conditional information or higher-order evidence.

To adapt an example from Weisberg (2015a): Suppose you are about to look at a wall in the Forbidden City. Let E be the proposition *The wall appears red to you*, R be the proposition *The wall is red*, and D be the proposition *You are a bad judge of your color experience*. Pre-theoretically, one might think whether one is good at introspecting one's color experiences is probabilistically independent of whether one's color experiences do in fact track the truth, i.e. it is rationally permissible that

$$C(R|D, E) = C(R|E).$$

Nevertheless, after you revise your credences in R on the basis of your visual appearance, you should lower your confidence if you learn in addition that you are a bad judge of your own color experiences, i.e.

$$f(E)(R|D) < f(E)(R).$$

But if $f(E)(E) = 1$, then this belief change is possible only if f is non-rigid, i.e. $f(E)(\cdot|E) \neq C(\cdot|E)$.¹⁰

But even if Rigidity is false, one might still question Self-Stability on the basis of the following principle:

Weak Evidentialism. It is always rationally permissible to plan to continue being uncertain of propositions that will not be entailed by one's total evidence, i.e.

$$\forall p, C(p) < 1 \& E(w) \not\subseteq p \Rightarrow \text{it is permissible that } f(w)(p) < 1.$$

One may think of Weak Evidentialism as a generalization of Weak Regularity (cf. Chapter 2).¹¹ Both are motivated by the thought that rationality cannot *require* absolute certainty in the absence of decisive evidence. If you learn that a fair die landed on an even number, then it is at least permissible for you to not jump to the conclusion that it landed on a six. Similarly, if your evidence will not entail information about what credence function you plan to adopt, then it is at least permissible for you plan to be uncertain about what credence function you plan to adopt. But Self-Stability says that all rational updating plans must be self-stable and thereby self-luminous. If you plan to adopt credence function P at w , then you should plan to be certain that one plans to adopt P at w , even if the information *You plan to adopt P* is not part of your evidence at w .

As mentioned in §3.2, I think Weak Evidentialism is false, despite its intuitive plausibility. Sometimes

¹⁰Admittedly, the Bayesian who thinks evidence is factive would probably insist that, if the wall does in fact look red to you, then you should not change your credence in R after learning that you are a bad judge of her color experience. You might think that you *should* have lower credence in R than you currently do, and to this extent this is a case of rational epistemic akrasia (cf. Williamson 2011, 2014; Lasonen-Aarnio 2014).

¹¹Recall that Weak Regularity says: if you are *a priori* indifferent between two disjoint possibilities, then learning their disjunction should not make you certain that one of them is true.

one is rationally required to *plan* to be certain of propositions that one is not certain of now and will not be entailed by one's total evidence. However, this does not mean that, in those circumstances, one is required to *be* certain of propositions that one *was* uncertain about and *are* not entailed by one's total evidence. These two claims may seem contradictory to each other, and it is my goal in the remainder of this chapter to make this conjunction less unpalatable.

Let's start with a somewhat fantastical thought experiment. Suppose the scientists have made a new breakthrough: they have invented a way of producing pills that would make us do certain things under certain conditions, with no side effects. For instance, there is a pill that would make one buy green onions if there are no red or yellow ones left in the grocery store; another pill that would make one buy no onions and walk home instead in that circumstance. There are also pills that would make one adopt different credence functions in response to different learning experiences. In fact, for every plan that we can conceive of, there is a pill that would achieve the same effect.

There is however one limitation to this new invention: the pills are effective, when and only when the pill-taker can respond to the relevant conditions. For instance, the pill that would make one go get a snack if and only if one is hungry is usually effective, whereas the pill that would make one go get a snack if and only if one's neighbour is hungry is usually not.

So in this imaginary world, whenever one is making a plan about what to do under various conditions, one has an alternative choice: pop the corresponding pill, and wait for the effect to kick in when the time comes.

Now consider *Unmarked Clock*, transplanted to this sci-fi world. Recall the case:

Unmarked Clock. You are trying to catch a bus that is supposed to arrive at 9:50am. You arrived at the bus stop, but there is no sign of the bus. Unfortunately your phone has run out of battery, and you've left your watch at home. So all you can do is squinting at a clock on a shopping mall nearby. To make matters worse, the clock is extremely austere – it only has a dial with no markings. Given your visual acuity, if the clock hand is pointing at time t , then you will get evidence that it is $t \pm 1\text{min}$, but not anything more precise.

Suppose you are about to take the pill – call it “the B-pill” – that would make you change your credence in accord with BCND when you look at the clock. In particular, if the time is t , then you will be confident that it is $t \pm 1\text{min}$, but you won't be able to rule out any time within that range with certainty.

At this point, Doctor Who shows up from nowhere, and gives you an offer,

Ah I see that you are taking the B-pill. Don't take that. Here I have something better for you.

Take the M-pill. It will make you infallibly confident about what the time is when you look at

the clock. And don't worry: if the B-pill is effective for you, then your degrees of belief must be very sensitive to what the time is – if the time were different, so would your credences. So the M-pill will be effective for you just as well.

Should you take the M-pill instead of the B-pill? I can't think of reasons why you may not. But Weak Evidentialism says that you may decline the offer. For by stipulation, if the time is t , then your evidence won't rule out the time being $t \pm 1$. So by Weak Evidentialism, it is permissible for you to plan not to be certain that the time is t when you look at the clock. Since taking a pill is effectively the same as making a plan that will be followed through, it would appear that you are as well permitted not to take the M-pill either, according to Weak Evidentialism.

The previous argument is an argument by analogy. Deliberating about which plan to follow is analogous to deliberating about which pill to take that would make you do what you plan to do. In *Unmarked Clock*, you are rationally required to take the M-pill. Therefore, in *Unmarked Clock*, you are rationally required to plan to be confident about what time it is when you look at the clock, despite the imprecision of your visual appearance.

Now there is one caveat: For the pill-popper, when the effects kick in at t , she has no choice but to manifest the action/belief that the pill would make her perform/adopt. But the planning agent can always in principle deviate from her plan. We may plan to do one thing, but end up doing something else, when the situation arises. Moreover, some of these revisions are not necessarily irrational. We are not invariably bound by our past choices; every day is a new day.

The same seems true of credal revision. It is one question how rational agents should plan to revise their credences, in response to each and every piece of evidence that she *will* get. It is another question what credence function a rational agent should adopt, in response to the evidence that she *has gotten*. The answers to these two questions converge, if and only if we assume that rational updating plans are binding. But it is unclear why they have to be. It may seem a good plan – maybe *the* rational plan – to be confident that it is 9:50am if and only if it is 9:50am. But when the evidence comes in, one has to work with what one got, and the evidence is indecisive.

So here is my final (though still tentative) diagnosis of *Opaque Three Prisoners*: It may be rationally permissible for Royo to *end up* being 1/2 confident that he will be freed, irrespective of what the guard tells him; it is just rationally impermissible for him to plan to adopt that credence *ex ante*. To this extent, cases like *Opaque Three Prisoners* are indeed similar to forgetting (though not in exactly the same way that Williamson had intended). It is rational to forget, but irrational to plan to forget. Similarly, Royo may rationally end up updating in accord with BCOND, though that should have never been the plan.

Chapter 5

Deference and Decisions

1 Introduction

The previous chapter proposes a new norm for rational plans: a rational plan should be **self-stable**, in the sense that a rational agent who endorses that plan should not plan to switch to a different plan, once she learns what the plan tells her to do. In particular, an updating plan f is self-stable if it equals itself conditional on its own recommendations, i.e.

$$f(w)(\cdot) = f(w)(\cdot | f = f(w)), \forall w \in W.$$

As noted there, with respect to updating plans, the principle of Self-Stability is entailed by the following principle:

Global Reflection. Rational agents should defer to their updating plans, i.e. for any prior C and updating plan f , (C, f) is rational only if

$$f(w)(\cdot) = C(\cdot | f = f(w)), \forall w \in W.$$

When C is held fixed, we call $f^*(w) = C(\cdot | f = f(w))$ **the reflective version of f** . We say C **defers to f** , or equivalently that f is **self-reflective**, if $f = f^*$. Note that the reflective version of an updating plan is always self-reflective, i.e. $(f^*)^* = f^*$.¹

We saw that the BCOND-plan can fail to be self-reflective (and self-stable). In *Opaque Three Prisoners*,

¹That is, $C(\cdot | f^* = f^*(w)) = C(\cdot | f = f(w))$ for all w . To see why, note that $f(v) = f(w) \Rightarrow f^*(v) = f^*(w)$. So $[f = f(w)] \subseteq [f^* = f^*(w)]$. Suppose $f(v) \neq f(w)$. Then $[f = f(v)] \neq [f = f(w)]$ and so $C(\cdot | f = f(v)) \neq C(\cdot | f = f(w))$ (assuming regularity). Thus $[f = f(w)] = [f^* = f^*(w)]$. So $f^*(w) = C(\cdot | f = f(w)) = C(\cdot | f^* = f^*(w)) = (f^*)^*(w)$.

the plan tells Royo to be $1/2$ confident that he will be pardoned, no matter what the guard will tell him by telepathy, though he should not adopt that credence now before he approaches the guard. He must wait until he reaches that bridge, and cross it with composure.

In the previous chapter, I argued against the rationality of BCOND in cases like *Opaque Three Prisoners* on the basis of an analogy. Deciding which plan to adopt is similar to deciding which pill to take that would be effective and make one act in accord with the plan. It is not always rational to take the pill that would make one update in accord with BCOND. So it is not always rational to plan to update in accord with BCOND.

Here is another way of stating the objection. The general question that we are interested in is: How should rational agents plan to update in response to different possible learning experiences? One way to approach that question is to treat it as a *cognitive decision problem*. Just as rationality requires taking the best pill that is effective, rationality requires adopting the best updating plan that is available. However, intuitively BCOND is not the optimal plan in cases like *Unmarked Clock* – the plan of being confident that the time is t if and only if the time is t seems much better, and is in a precise sense just as available as the BCOND-plan.

The main goal of this final chapter is to precisify and generalize this insight by proving two technical results:

1. A non-self-reflective updating plan is no better than its reflective version;
2. Between two self-reflective updating plans, the one that is more sensitive is the better.

Taken together, they entail the following theorem:

Decision-theoretically speaking, the best updating plan is always the one that recommends conditionalizing on the true element of the partition to which one is most sensitive.

In Meehan and Zhang (manuscript), we call this principle “sensitivity conditionalization,” or S-COND for short. S-COND disagrees with BCOND in all but exceptional circumstances. To this extent, what appears to be a conflict between Global Reflection and Conditionalization turns out to be a more fundamental conflict between Bayesianism on the one hand, and cognitive decision theory on the other.

Plan. §2 sets the stage and introduces some terminologies (this section can be skipped and referred back to when necessary). §3-4 prove the two main results. §5 discusses their general philosophical upshots.

2 Setup

As before, I’ll assume that rational agents have degrees of belief that can be represented by a real-valued partial function C over the powerset of some space W . Unless otherwise specified, I’ll assume that W is finite

and the agent's prior C is a unary, regular probability function that is defined over $\mathcal{P}(W)$. In particular, for any $E \subseteq W$, $C(.|E)$ is well-defined and given by the ratio formula.

Recall that an **evidential situation** is representable by a function E , which assigns a world $w \in W$ to a proposition $E(w)$, interpreted as the agent's total evidence at world w . Fix an evidential situation E . For every $w \in W$, the set $[E = E] = \{v \in W : E(v) = E\}$ represents the proposition *E is your (i.e. the agent's) total evidence*. We say an evidential situation is **factive** if $w \in E(w)$ for all $w \in W$ (or equivalently $[E = E] \subset E(w)$); it is **luminous** if $E(w) \subset [E = E(w)]$.

Some think all possible evidential situations are factive and luminous. Such theorists deny the cogency of cases like *Opaque Three Prisoners*. For the present purposes, I shall assume that non-luminous evidential situations are possible. However, as I discuss in §5, the main thesis of this chapter does not depend on that assumption; the BCOND-plan may be sub-optimal even in luminous evidential situations. To this extent, the tension between Bayesianism and cognitive decision theory stands orthogonal to the metaphysical debate about Evidential Luminosity.

Now let's change gears to updating plans. As noted in Chapter 2, we can think of a plan as a family of conditionals whose antecedents form a partition. Without loss of generality, we can also assume that the consequents of these conditionals are all distinct. If a plan says "if C_1 then A " and "if C_2 then A ," then it can be equated with a plan that says "if C_1 or C_2 then A ." Fix a plan f . We say a proposition E is an **antecedent of f** if it is logically equivalent to an antecedent of a conditional that constitutes f . By assumption, the set of antecedents of f is equivalent to the partition $\mathcal{E}_f = \{[f = f(w)] : w \in W\}$.

Recall that a plan f' is **at least as available as f** just in case $f(w) = f(v) \Rightarrow f'(w) = f'(v)$. Equivalently, f' is at least as available as f if $\mathcal{E}_f$ is a refinement of $\mathcal{E}_{f'}$.² As mentioned in Chapter 4, one way to think of availability is that it (roughly) tracks followability: an updating plan is more available if it has weaker antecedents and thereby is less demanding about what conditions the agent would need to respond to in order to follow the plan. For purely aesthetic reasons, I shall sometimes speak of *sensitivity* instead of *availability*: f' is at least as available as f just in case it is as most as sensitive as f . Let $\preceq$ denote the "being at most as available as" or "being at least as sensitive as"-relation. Note that $\preceq$ is a partial order, i.e. it is reflexive, antisymmetric and transitive. Moreover, the reflective version of an updating plan (with respect to a fixed prior) is always at least as available as itself, i.e. $f \preceq f^*$.

Besides comparing the availability of different updating plans, we may also assess the availability of one specific updating plan with respect to an evidential situation. We say an updating plan f is **available to an evidential situation E** (or **E -available**) just in case it does not give different recommendations at two states where the agent's evidence is the same, i.e. $E(w) = E(v) \Rightarrow f(w) = f(v)$. Equivalently, f is E -available

²A $\mathcal{E}$ refines another partition $\mathcal{E}'$ just in case every element of $\mathcal{E}'$ is a union of elements of $\mathcal{E}$.

just in case the partition $\mathcal{E} = \{[E = E(w)] : w \in W\}$ is a refinement of $\mathcal{E}$. More generally, if $\mathcal{E}$ is a partition, then f is $\mathcal{E}$ -**available** if it is constant over any particular element of $\mathcal{E}$, i.e. $w, v \in E \in \mathcal{E} \Rightarrow f(w) = f(v)$ (or equivalently $\mathcal{E}$ is a refinement of $\mathcal{E}_f$).

For any evidential situation E , the BCOND-plan with respect to E is always E -available. But in general it would not be the only updating plan that is E -available. Here is another example. Fix C and E . Let f be given by

$$f : w \mapsto C(.|E = E(w)).$$

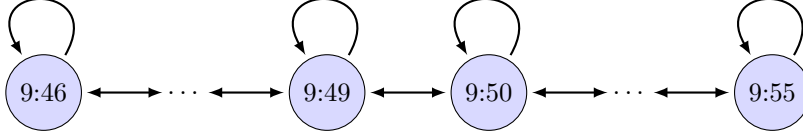
Intuitively, the plan says: if E is your total evidence, then update by conditionalizing on the proposition E *is your total evidence*. Following Das (forthcoming), I call this plan *meta-conditionalization*, or META-COND for short. One interesting thing to note is that, for any prior and evidential situation, the META-COND-plan *is* the reflective version of the BCOND-plan. By our previous observation, it follows that the META-COND-plan is always at least as available as the BCOND-plan.

One last piece of notation: For a probabilistic credence function C and a partition $\mathcal{E}$, I'll let $C(.|\mathcal{E})$ represent the plan of updating by conditionalizing on the true element of $\mathcal{E}$, i.e. for every w , $C(.|\mathcal{E})(w) = C(.|E_w)$, where $E_w \in \mathcal{E}$ is the unique partition element containing w . It is not hard to check that, for any C and $\mathcal{E}$, $C(.|\mathcal{E})$ is $\mathcal{E}$ -available and self-reflective.

Before proceeding: Throughout this chapter, I shall be using *Unmarked Clock* as my working example. So it would be helpful to remind ourselves of the case:

Unmarked Clock. You are trying to catch a bus that is supposed to arrive at 9:50. You arrive at the bus stop, but there is no sign of the bus. Unfortunately your phone has run out of battery, and you've left your watch at home. So all you can do is squinting at a clock on a shopping mall nearby. You know that the clock is accurate up to minutes. But unfortunately, it is extremely austere – it only has a dial with no markings.

For simplicity, say you know that it is definitely past 9:45 and before 9:56, and your credence is uniformly distributed across that range (coarse-grained to minutes). For every time t , let $t \pm 1$ denote the time 1 min before/after t . Suppose that, if the actual time is t , then the strongest proposition you can learn by squinting at the clock is that the actual time is $t \pm 1$. Importantly, when you learn that the actual time is $t \pm 1$, you do not learn that *that* is what you've learnt (for otherwise you would be able to deduce that the actual time is t). That is, your evidential situation is non-luminous (cf. Chapter 4).



Let f be the BCOND-plan given your prior and evidential situation in *Unmarked Clock*. So for every t between 9:47 and 9:53 (inclusive), f recommends you to assign a credence of $1/3$ to $t - 1, t$ and $t + 1$, whereas if the actual time is 9:46 (or 9:55), then f says you should split your credence between 9:46 and 9:47 (or 9:55 and 9:54).

Note that, in this model, you learn something slightly different by looking at the clock at different times. For instance, if the actual time is 9:50, then you learn that it is not 9:48, but you can't rule out it being 9:49, whereas if it is 9:51, then you can rule out it being either 9:48 or 9:49. Formally, this means $[E = E(w)] = \{w\}$. So the META-COND-plan (or equivalently f^*) is the omniscient plan: if the actual time is t , then you should be fully confident that it is t .

3 First Result

The goal of this section is to prove my first result: a non-self-reflective updating plan is no better than its reflective version. In more precise terms:

For any prior C and updating plan f , if C does not defer to f (i.e. f is not self-reflective), then the reflective version of f is strictly better than f in terms of (i) accuracy and (ii) practical utility.

Let me start with accuracy.

3.1 Accuracy

Recall that an accuracy scoring rule is a function that assigns real numbers to credence-world pairs:

$$\mathcal{A} : (C, w) \mapsto \mathcal{A}(C, w) \in \mathbb{R}$$

For each pair (C, w) , $\mathcal{A}(C, w)$ measures how accurate C is at w .

Intuitively, if we think of being accurate as “getting things right,” i.e. assigning high credence to truths and low credence to falsehoods, then the most accurate credence function at w is the omniscient credence function v_w , which assigns 1 to all and only propositions that are true at w . That is, for any C ,

$$\mathcal{A}(C, w) \leq \mathcal{A}(v_w, w)$$

with equality if and only if $C = v_w$.

Clearly META-COND is better than BCOND accuracy-wise in *Unmarked Clock*. Just as no credence function could be more accurate than the omniscient credence function, no updating plan could be more accurate than the omniscient plan.

There are two ways of making this intuition more precise. The first is by looking at the plans' respective *expected accuracy*. Recall that the expected accuracy of a credence function C' relative to C is given by the C -mean accuracy score of C' :

$$\mathbb{E}_C[\mathcal{A}(C')] = \sum_w C(w) \mathcal{A}(C', w).$$

Similarly, we can calculate the expected accuracy of an updating plan by the C -mean accuracy of the credence functions recommended by f :

$$\mathbb{E}_C[\mathcal{A}(f)] = \sum_w C(w) \mathcal{A}(f(w), w).$$

It's not hard to see that $\mathbb{E}_C[\mathcal{A}(f^*)]$ is strictly greater than $\mathbb{E}_C[\mathcal{A}(f)]$ in *Unmarked Clock*. To generalize this result, we need one more assumption about accuracy scoring rules:

Strict Propriety. An accuracy scoring rule does not make rational credence functions modest. In particular, it does not make any probabilistic credence function immodest, i.e. if $\mathcal{A}$ is an accuracy scoring rule, then for any probability function P and credence function Q ,

$$\mathbb{E}_P[\mathcal{A}(P)] \geq \mathbb{E}_P[\mathcal{A}(Q)]$$

with equality if and only if $P = Q$.

As noted in Chapter 4, Strict Propriety follows from two intuitively plausible thoughts: (i) rational credence functions should be immodest, and (ii) every probabilistically coherent function can in principle be rational given the right circumstances. It turns out that, assuming Strict Propriety, f^* always surpasses f in terms of expected accuracy.

Fact 1. Fix a regular prior C and a strictly proper accuracy score $\mathcal{A}$. For any updating plan f ,

$$\mathbb{E}_C[\mathcal{A}(f^*)] \geq \mathbb{E}_C[\mathcal{A}(f)].$$

with equality if and only if C defers to f (i.e. $f = f^*$).

An immediate corollary of Fact 1 is that BCOND does not maximize expected accuracy if it is non-self-reflective and its reflective version (i.e. META-COND) is also in the pool. Since self-reflectiveness implies

self-stability, and BCOND is self-stable if and only if the evidential situation is luminous (cf. Chapter 4), it follows that BCOND does not maximize expected accuracy in non-luminous evidential situations (provided that META-COND is also available).

Aside: Schoenfield (2017) proves a stronger fact: in any evidential situation E , not only does META-COND dominate BCOND in terms of expected accuracy, but it dominates any other E -available updating plan in that respect as well,³ which she in turn generalizes to the following fact:

Super-Duper (SD) Generalized CondMax. Consider any partition of propositions $\mathcal{E} = \{E_i\}_{i \in I}$ over a set of states W . Let f be a function from $\mathcal{E}$ to credence functions with the intended interpretation that an agent adopts $f(E)$ whenever E obtains. The f that maximizes expected accuracy is the one that assigns to each E the credence function that results from conditionalizing on E .⁴

In the §4, I shall show that both of Schoenfield’s result can indeed be proved as a corollary of my two results, of which Fact 1 is a part.

We have seen one way in which f^* is better than f : it has higher expected accuracy with respect to the background prior. However, in *Unmarked Clock*, there seems to be a more direct sense in which the META-COND-plan f^* is more accurate than the BCOND-plan f : f^* strongly accuracy-dominates f , i.e.

$$\forall w \in W, \mathcal{A}(f^*(w), w) > \mathcal{A}(f(w), w).$$

A natural question to ask is whether this property also generalizes to other pairs of updating plans and their reflective versions.

Unfortunately, it does not. For instance, fix $W = \{a, b\}$. Let C be the uniform probability distribution. Consider the updating plan f that recommends adopting the omniscient credence function at world a no matter what comes, i.e. $f(w) = v_a$ for all w . Then $f^*(w) = C(.|f = f(w)) = C$. Since $C \neq v_a$ and v_a has

³Recall that E -availability means that $E(w) = E(v) \Rightarrow f(w) = f(v)$, or in English: the plan does not recommend different credence functions at states where the agent has the same evidence.

⁴Strictly speaking Schoenfield’s formulation seems ambiguous between two readings:

1. For any partition $\mathcal{E}$, among the updating plans that can be interpreted as a family of conditionals with antecedents from $\mathcal{E}$ (i.e. if E then...), the updating plan that maximizes expected accuracy is $C(.|\mathcal{E})$;
2. For any set of updating plans Φ and any partition $\mathcal{E}$, if every updating plan in Φ can be interpreted as a family of conditionals with antecedents from $\mathcal{E}$, then the one that maximizes expected accuracy among members of Φ is $C(.|\mathcal{E})$.

The second reading is clearly false, since every updating plan can be interpreted as a family of conditionals with antecedents from the finest partition $\mathcal{E} = \{\{w\} : w \in W\}$. But in this case $C(.|\mathcal{E})$ would be the omniscient plan, which may not even be a member of Φ . So I assume that Schoenfield intended the first reading. This also brings out an important contrast between our results. Note that an updating plan f only maximizes expected accuracy with respect to a set Φ of updating plans fixed in the background. Our two results differ with respect to how the set Φ is chosen. In SD Generalized CondMax, Φ is determined by a partition; in my result, Φ is picked at random, and the partition is recovered from Φ (assuming that Φ is closed under reflection).

the highest accuracy score at world a , we have

$$\mathcal{A}(f(a), a) = \mathcal{A}(v_a, a) > \mathcal{A}(C, a) = \mathcal{A}(f^*(a), a).$$

So in this case f^* does not strongly accuracy-dominate f .

Nevertheless, there is a different sense in which f – or more precisely (C, f) taken together as a pair – is accuracy-dominated whereas (C, f^*) is not. For concreteness, let's say $\mathcal{A}$ is given by the Brier score, i.e.

$$\mathcal{A}(C, w) = - \sum_{p \subseteq W} (C(p) - v_w(p))^2.$$

In the above example, if one starts out with prior C and updates in accord with f , then the accuracy score one would accrue at a world w at each stage is given by the following table:

	a	b
$\mathcal{A}(C, w)$	-0.5	-0.5
$\mathcal{A}(f(w), w)$	0	-2
Total	-0.5	-2.5.

On the other hand, suppose one starts out with $C' = (0.7, 0.3)$ instead, and updates in accord with the dummy plan f' that tells one to retain one's prior, i.e. $f'(w) = C'$ for all $w \in W$. Then the total accuracy score one would accrue at each world is given by

	a	b
$\mathcal{A}(C', w)$	-0.09	-0.98
$\mathcal{A}(f'(w), w)$	-0.09	-0.98
Total	-0.18	-1.96.

Notably, although C' does not strongly accuracy-dominate C , nor does f' strongly accuracy-dominate f , the total accuracy score of C' and f' combined does strongly accuracy-dominate that of C and f combined. Moreover, like f , f' is also a plan that recommends adopting one credence function no matter what comes. So in my terminology, f' is at least as available as f : $f(w) = f(v) \Rightarrow f'(w) = f'(v)$.

By contrast, one can show that (C, f^*) cannot be accuracy-dominated in a similar fashion. That is, there are no C' and updating plan f' that is at least as available as f such that

$$\mathcal{A}(C, w) + \mathcal{A}(f^*(w), w) < \mathcal{A}(C', w) + \mathcal{A}(f'(w), w), \forall w \in W.^5$$

⁵To spell out the reasoning in more detail: Note that $\mathcal{A}(C, w) + \mathcal{A}(f(w), w) = 2\mathcal{A}(C, w)$. Similarly, since f' is at least as

This asymmetry turns out to generalize to all updating plans and their reflective versions (the following result is proven as a corollary of Briggs and Pettigrew, forthcoming):

Fact 2. *Fix a strictly proper accuracy score $\mathcal{A}$.*

1. *If C defers to f , then (C, f) is not strongly accuracy-dominated by another pair (C', f') where f' is at least as available as f .*
2. *Given additional assumptions about $\mathcal{A}$ (e.g. $\mathcal{A}$ is the Brier score), if C does not defer to f , then (C, f) is strongly accuracy-dominated by another pair (C', f') where C' defers to f' and f' is at least as available as f .*

To take stock: We have seen two respects in which the reflective version of an updating plan is better in terms of accuracy: (i) it has higher expected accuracy, and (ii) it avoids accuracy domination.

Some think an updating plan is rational for an agent just in case it best promotes accuracy with respect to the agent’s prior and constraints (e.g. Pettigrew 2016b). Some (myself included) think there are other dimensions to the rational assessment of updating plans as well. One of them is practical deliberation. We deliberate on the basis of our credences. Intuitively, an updating plan is better if updating in accord with that plan helps us make better decisions in general. The goal of the next subsection is to make this pragmatic “betterness” relation more precise, and proves two results parallel to Fact 1 and 2 above.

3.2 Decision-making

Let’s go back to *Unmarked Clock*. Same set-up as before, except that this time you have a company at the bus stop. In fact, the guy is a gambler. He turns around to you with the following offer:⁶

Before you check the time, let’s make a deal. Here I have two bets, both cost \$10. The first one pays you back \$20 if the dial is pointing at an odd number, the second one pays the same amount if the dial is pointing at an even number. Now look at the clock, and then take either bet as you wish.

available as f , it too must also be a constant function. Say $f(w) = P$ for all w . By Strict Propriety:

$$2\mathbb{E}_C[\mathcal{A}(C)] \geq \mathbb{E}_C[\mathcal{A}(C')] + \mathbb{E}_C[\mathcal{A}(P)]$$

So

$$\begin{aligned} & C(\{a\})(\mathcal{A}(C, a) + \mathcal{A}(f^*(a), a)) + C(\{b\})(\mathcal{A}(C, b) + \mathcal{A}(f^*(b), b)) \\ & \geq C(\{a\})(\mathcal{A}(C', a) + \mathcal{A}(f'(a), a)) + C(\{b\})(\mathcal{A}(C', b) + \mathcal{A}(f'(b), b)) \end{aligned}$$

which means that the total accuracy score (C, f) cannot be strongly dominated by that of (C', f') .

⁶Ahmed and Salow (2019) consider a similar case. It may be worth noting that the question I am asking here is different from theirs. They are interested in the question “When should you look at the clock, holding fixed your updating plan?” whereas I am interested in the question “Given that you are going to look at the clock, how should you plan to update?”

If you plan to update in accord with BCOND, then you should know better not to take the deal. For you know *ex ante* that if the dial is pointing at an odd/even number between 9:47 and 9:54, then you will be 2/3 confident that it is pointing at an even/odd number, which means you will almost always find the losing bet attractive. On the other hand, if you plan to update in accord with META-COND, then you can rest reassured that your future self will only take the bet that wins money (bracketing possibilities where you make a mistake or prefer giving charity to strangers). So in *Unmarked Clock*, the META-COND-plan is better than the BCOND-plan in not only alethic but also pragmatic terms; it helps you choose wisely. (As we shall see, it is not a coincidence that these two perspectives coincide – the formal frameworks are almost isomorphic, only endowed with different interpretations.)

Again, there are two ways of making this intuitive observation more precise. First, we may look at “the worst” that could happen if you update in accord with either plan. Before we assume that the gambler only asks you to choose between two bets after you look at the clock. But suppose he also offers you a third bet, to be accepted or declined before you look at the clock:

Pay me \$16, and if the time is not 9:46 or 9:55, then I’ll give you \$21.

By assumption, you are 0.8 confident that the time is not 9:46 or 9:55. So from your perspective, the expected payoff of this bet is

$$0.8(\$5) - 0.2(\$16) = \$0.8$$

which is not bad. So you might as well take it. However, if you in addition look at the clock, update in accord with BCOND and deliberate about the gambler’s first deal on the basis of the credences prescribed by the plan, then you would end up losing at least \$5, no matter what the time is.⁷

One way of characterizing your predicament above is that you are *Dutch-booked*. The word “Dutch book” has however been used to refer to many different things. Here is what I mean by this term in the present context. Imagine an agent with prior C and updating plan f . Suppose that, before she updates, she is offered two sets of bets, $\mathcal{O}_1$ and $\mathcal{O}_2$, both of which contain the null bet. The agent is asked to take one from each, before and after she updates respectively. Her choices are represented by a pair (X, S) – call it a **choosing strategy** – where $X \in \mathcal{O}_1$ is the bet that she will take from $\mathcal{O}_1$, and S is a function that maps each world w to the bet $S(w)$ that she will take from $\mathcal{O}_2$ at world w . We assume in addition that the agent is disposed to choose similarly whenever her credence functions are the same, i.e. $f(w) = f(v) \Rightarrow S(w) = S(v)$.

We say (X, S) is **acceptable** (with respect to (C, f) and $(\mathcal{O}_1, \mathcal{O}_2)$) if each bet has non-negative expected payoff with respect to the agent’s credence function at the world and the time when it is accepted. A **Dutch**

⁷If the time is not 9:46 or 9:55, then you would take the losing bet that costs you \$10 after you look at the clock, which is not fully compensated by your \$5 from the first bet. On the other hand, if the time is 9:46 or 9:55, then the most you can gain from your second bet is \$10, which again won’t compensate your loss from the first bet, which is \$16.

book against (C, f) is then a pair of bets $(\mathcal{O}_1, \mathcal{O}_2)$ such that there is some acceptable (X, S) with respect to $(\mathcal{O}_1, \mathcal{O}_2)$ and (C, f) that guarantees the agent a sure loss in the following sense:

$$\forall w \in W, X(w) + S(w)(w) < 0.$$

Fact 3. *A pair (C, f) is Dutch-bookable if and only if C does not defer to f .*

One immediate corollary of Fact 3 is that the BCOND-plan is Dutch-bookable (in combination with the agent’s prior) whenever it disagrees with the META-COND-plan, which is not Dutch-bookable. Das (forthcoming) proves a stronger fact: in a fixed evidential situation, the META-COND-plan is the only updating plan that is non-Dutch-bookable. Again, as I show in §4, there is a precise sense in which Fact 3 together with another result jointly entails (a version of) Das’ result as a corollary.

Before moving on, let me make a remark about my definition of Dutch book (this can be skipped without loss of continuity).

Remark 1. Admittedly, my account of Dutch book is slightly unconventional. On the orthodox account (cf. Lewis 1999; Das, forthcoming), a diachronic Dutch book consists of a sequence of individual bets, to be offered by the bookie to the agent before and after she updates. The bookie is assumed to know the agent’s credence and updating plan, but nothing else. The agent is Dutch-booked if the bets are individually acceptable but jointly guarantee her a sure loss. It is not hard to see that, if we pool all the possible pre-updating and post-updating bets that the bookie could potentially offer to the agent into two sets, then this is equivalent to my definition. The only difference is that, on my account, it is the agent rather than the bookie who chooses which bet is to be taken (“offered”) from the respective pool.

I prefer this alternative formulation for three reasons. First, it is more closely related to an alternative, stronger definition of Dutch book that we will discuss shortly. Second, if the bookie is the one who determines which bet will be offered after the agent updates, then it appears that he must have some epistemic access to the agent’s post-learning credences, which the agent may lack herself (though see Gallow (2019)). But then Dutch-bookability may not seem so problematic after all. It is not one’s fault to be exploitable by someone more knowledgeable; ignorance is not a rational sin. On the other hand, I think it is more plausible that the agent may make the choice on the basis of her credences, even if she does not know what they are; one may follow one’s heart without knowing where it is going.

Third, most importantly, one might think that, once the bookie makes the offer, the agent should deliberate not on the basis of her post-learning credences, but her post-learning credences *conditional on the bookie having made the offer to her*. Formally, if the agent updates in accord with an updating plan f , and

the bet that the bookie will offer to the agent at w is $B(w)$, then the orthodox account has it that $B(w)$ only needs to be acceptable relative to $f(w)$, whereas it would appear that, once the bookie makes the offer, the agent should deliberate on the basis of not $f(w)$ but $f(w)(\cdot|B = B(w))$. If we impose this different constraint, then it is not true that agents are Dutch-bookable (in this new sense) if they don't defer to their updating plans. By contrast, I think it is again more plausible that the agent may choose without knowing what she is about to choose; we sometimes surprise even ourselves. Of course, she may regret instantly as soon as she takes the bet, but a deal is a deal.⁸

There is however one wrinkle with this result. Before I defined a Dutch book to be a set of bets for which some *acceptable* choosing strategy ends up losing the agent money no matter what comes. But one might think, in order for Dutch-bookability to reveal a rational defect, it is insufficient that the choosing strategy is merely *acceptable*; rather, it should be *optimal*. If you are offered \$10 for free and you value money, then you should take it rather than walk away, even though walking away is an “acceptable” option in the sense that it won't lose you anything. The same is true of your future self. For every w , your future self should accept $S(w)$ at w only if you regard it as the best, or one among the best, of all options in $\mathcal{O}_2$. If it is not, then the fact that (X, S) incurs you a sure loss may not be a sign of any rational defects of your credences and/or updating plans; you are simply making bad (though acceptable) choices.

Say a choosing strategy (X, S) (with respect to (C, f) and $(\mathcal{O}_1, \mathcal{O}_2)$) is **optimal** if X maximizes expected utility with respect to C and $S(w)$ maximizes expected utility with respect to $f(w)$ for every w . We say $(\mathcal{O}_1, \mathcal{O}_2)$ is a **strong Dutch book for (C, f)** if (C, f) licenses an optimal choosing strategy that incurs the agent a sure loss in the above sense.

As we saw above, in *Unmarked Clock*, the BCOND-plan is not only Dutch-bookable but also strongly Dutch-bookable; on the other hand, since the META-COND-plan is not Dutch-bookable, *a fortiori* it is not strongly Dutch-bookable. So one natural question to ask is: Does this result generalize to all updating plans and their reflective versions? Is (C, f) strongly Dutch-bookable, whenever C fails to defer to f ?

Unfortunately, the answer is negative. Here is a simple counter-example: recall *Red Wall*. Formally, the case can be modelled as follows: Let $W = \{a, b\}$; a represents the “good case” where the wall is red, and b represents the “bad case” where the wall is white but appears red. Let C be the uniform probability

⁸Admittedly, (as Adam Elga pointed out to me in conversation) there is a trade-off here. While my notion of Dutch-bookability does away with potentially problematic knowledge assumptions, it also seems normatively weaker. For a Dutch book in my sense no longer *guarantees* a sure loss. When the bookie offers the agent the two sets of bets, in general he would not know whether the agent would take bets in a way that would guarantee her a sure loss. Similarly, from the agent's perspective, being Dutch-bookable in my sense only means that she opens herself up to the *possibility* of voluntarily choosing in a way that would incur her a sure loss (in fact it may not even be possible that she would choose in this way, as discussed below).

distribution over W . The BCOND-plan corresponds to the function

$$f(a) = (1, 0), \quad f(b) = (1/2, 1/2).$$

Fact 4. *In Red Wall, (C, f) is Dutch-bookable but not strongly Dutch-bookable.*

For the sake of illustration, let's work through the case of *Red Wall*. Here is one way in which a gambler who knows your prior and updating plan could Dutch book you in my stipulated sense. Before you turn at the corner, they approach you with the following two bets:

Bet 1: buy a bet at \$5 that pays \$15 if the wall is white and \$0 otherwise.

Bet 2: sell a bet for \$10 that pays \$20 if the wall is white and \$0 otherwise.

You are to decide whether to take Bet 1 now, and whether to take Bet 2 after you make the turn. An acceptable strategy that would get you into trouble is the following:

$$X = \text{Bet 1}, S(a) = \text{Null}, S(b) = \text{Bet 2}.$$

	a	b
X	-\$5	\$15
S	\$0	-\$20
net	-\$5	-\$5

But of course (X, S) is not the optimal strategy. In particular, if the wall is red, then you would be certain that it is red, which you means should take Bet 2, *contra* what S says. Declining Bet 2 in this case would be like choosing to walk away when you are offered \$10 for free; it is acceptable, but not optimal.

More generally, Dorst et al. (manuscript) prove the following result:

Fact 5. *A pair (C, f) is not strongly Dutch-bookable if and only if (C, f) satisfies the following property*

Total Trust. *For any bet X and real number t , $\mathbb{E}_C[X | \mathbb{E}_f[X] \geq t] \geq t$.*

What shall we make of this? One tempting thought is that f^* is not necessarily better than f from the practical point of view after all (equivalently, deference to updating plan is at least not a requirement of practical rationality; this appears to be the lesson drawn by Dorst et al.). While I am sympathetic to the conclusion – I am only attracted to Self-Stability, which is strictly weaker than Global Reflection – I also think this inference may be a little too quick. Even if strong Dutch-bookability is more worrying than mere

Dutch-bookability, the latter may still seem like a defect that is better to be avoided; a small wound is a wound nonetheless.

More importantly, even if we assume that rational agents only take what is optimal, there is still a precise sense in which f^* always recommends a better *optimal* choosing strategy than f . This brings us to the second way of comparing the practical utilities of updating plans: instead of looking at “the worst” that could happen if one updates in accord with a plan, we look at “the best” that one can achieve, if one updates as the plan prescribes. More precisely: Let $\mathcal{O}$ be a finite set of bets. Suppose that, at each world w , the agent will choose a bet from $\mathcal{O}$ that maximizes her expected utility, after she updates in accord with her plan. Let S_f be the function that maps each world w to the bet that she will choose. As before, we assume that the agent is disposed to take the same bets at any two worlds where her credences are the same, i.e. $f(w) = f(v) \Rightarrow S_f(w) = S_f(v)$. Let $\mathbb{E}_C[S_f]$ be the expected utility of the diagonal function $w \mapsto S_f(w)(w)$, i.e.

$$\mathbb{E}_C[S_f] = \sum_w C(w) S_f(w)(w).$$

Intuitively, $\mathbb{E}_C[S_f]$ represents the agent’s pre-updating expectation of the net profit that she will accrue by choosing *optimally* in the decision problem that she will face, after having updated in accord with her plan.

Fact 6. *Fix C and a set of bets $\mathcal{O}$. Then for any updating plan f ,*

$$\mathbb{E}_C[S_{f^*}] \geq \mathbb{E}_C[S_f].$$

Moreover, if $\mathcal{O}$ is large enough such that $S_f \neq S_{f^*}$, then the equality is strict whenever $f \neq f^*$.

So, although neither the BCOND-plan nor the META-COND-plan is strongly Dutch-bookable in *Red Wall*, *ex ante*, you would still think that you will make better choices if you update in accord with META-COND rather than BCOND. In particular, you would know better not to take Bet 2 if the wall is white. (Formally speaking, (C, f) is strongly Dutch-bookable whenever $\mathbb{E}_C[S_f] < 0$. So what Fact 4 shows is that, even in cases where $\mathbb{E}_C[S_{f^*}]$ and $\mathbb{E}_C[S_f]$ are both non-negative, the former still trumps the latter unless $f = f^*$.)

To summarize what we have learnt so far:

Result I. For any pair (C, f) , TFAE:

1. $f \neq f^*$ (i.e. C does not defer to f);
2. $\mathbb{E}_C[\mathcal{A}(f^*)] > \mathbb{E}_C[\mathcal{A}(f)]$;
3. (C, f) is accuracy-dominated;
4. (C, f) is Dutch-bookable;

$$5. \mathbb{E}_C[S_{f^*}] > \mathbb{E}_C[S_f].$$

Result I tells us that, in a precise sense (or rather four precise senses), the reflective version of a non-self-reflective updating plan is strictly better. In particular, when they disagree, the META-COND-plan is always better than the BCOND-plan.

However, while Result I gives us a way of comparing updating plans with their reflective versions, it is silent on how to rank amongst the updating plans that are already self-reflective. Yet intuitively some self-reflective updating plans are clearly better than others. Consider the omniscient plan vs. the dummy plan; one tells the agent to believe the truth and nothing but the truth, the other tells the agent to retain her prior no matter what comes. Intuitively, the omniscient plan is better than the dummy plan both accuracy-wise and practically speaking, and we would like a way of making that intuition precise.

This brings me to my second result.

4 Second Result

It turns out that *sensitivity* is the key. Recall that an updating plan f is at least as sensitive as f' (or equivalently f' is at least as available as f) just in case $f(w) = f(v) \Rightarrow f'(w) = f'(v)$. Clearly the omniscient plan is more sensitive than the dummy plan. In fact the two represent the two ends of the sensitivity-spectrum, with the META-COND-plan sitting somewhere in between, depending on the respective evidential situation. Intuitively, provided that an updating plan is self-reflective, the more sensitive it is, i.e. the closer it is to the omniscient plan on the sensitivity-spectrum, the more accurate it should be.

This is indeed true, as witnessed by the following fact.

Fact 7. *Fix a regular prior C and a strictly proper accuracy score $\mathcal{A}$. If C defers to both f and f' , then*

$$\mathbb{E}_C[\mathcal{A}(f)] > \mathbb{E}_C[\mathcal{A}(f')] \Leftrightarrow f \prec f'.$$

Moreover, the phenomenon is universal: f is better than f' in each of the other three respects if and only if $f \prec f'$. However, stating these results precisely would involve going through extra technical details that may be more distraction than they are worth. So I shall refer the reader to the Appendix instead, and include here an informal statement of the result:

Result II. Fix a sample space W and a regular prior C . For any self-reflective f, f' , TFAE

1. $f \prec f'$;
2. $\mathbb{E}_C[\mathcal{A}(f)] > \mathbb{E}_C[\mathcal{A}(f')]$;

3. (C, f') is “ $\mathcal{E}_f$ -accuracy-dominated;”
4. (C, f') is “ $\mathcal{E}_f$ -Dutch-bookable;”
5. f recommends an optimal “ $\mathcal{E}_f$ -choosing strategy” with lower expected utility.

Here is a way of summarizing the upshot. Let Φ be an arbitrary set of updating plans (defined over a fixed sample space W). Let $\Phi^* = \{f^* : f \in \Phi\}$ – the set of reflective versions of members of Φ . We say Φ is **closed under reflection** if it includes the reflective version of every one of its members, i.e. $f \in \Phi \Rightarrow f^* \in \Phi$, or equivalently that $\Phi^* \subseteq \Phi$. Result I says that, for any subset of updating plans Φ closed under reflection, the best one must be self-reflective (i.e. it must be a member of Φ^*); Result II says that, with respect to the self-reflective updating plans, an updating plan is better if it is more sensitive. Putting these two pieces together, we arrive at the following general result:

Theorem 1. *For any Φ closed under reflection, the best updating plan of Φ is the most sensitive, self-reflective plan, namely $C(.|\mathcal{E}_\Phi)$ where $\mathcal{E}_\Phi$ is the coarsest partition that refines $\mathcal{E}_f$ for all $f \in \Phi$ (i.e. $\mathcal{E}_\Phi = \bigwedge_{f \in \Phi} \mathcal{E}_f$).*

If Φ is the set of **E**-available plans for some factive and luminous **E** (which entails that $\Phi = \Phi^*$), then we recover the classic results that **BCOND** is optimal in each of the four respects (Lewis 1999; Brown 1976; Greaves and Wallace 2006; Briggs and Pettigrew, forthcoming). If Φ is the set of **E**-available plans for an arbitrary evidential situation **E** (with no restrictions), then we recover Schoenfelds’ and Das’s results that **META-COND** maximizes expected accuracy and avoids Dutch books; it also follows that **META-COND** is also optimal with respect to pragmatic utility and accuracy domination (the latter answers a question raised in Nielsen (forthcoming)). Lastly, when Φ is the set of $\mathcal{E}$ -available plans, we get Schoenfeld’s second result (i.e. the updating plan that maximizes expected accuracy is $C(.|\mathcal{E})$), as well as its analogous versions for the other three measures of goodness of updating plans.

5 Sensitivity and rationality

One philosophical upshot of Theorem 1 is this: If we think of Φ as the set of updating plans that the agent can follow, and followability presupposes sensitivity, then we can think of $\mathcal{E}_\Phi$ as the set of most fine-grained conditions to which the agent is sensitive. Theorem 1 then suggests the following normative connection between sensitivity and luminosity:

Sensitivity Implies Luminosity (SL). If you are sensitive to whether E obtains, then you should plan to be infallibly certain about whether E obtains. More precisely: if Φ is the set of available updating plans

(closed under reflection) and $E \in \mathcal{E}$ for some $\mathcal{E}$ coarser than $\mathcal{E}_\Phi$, then $f \in \Phi$ is rational only if

$$f(w)(E) = 1 \text{ if } w \in E \text{ and } 0 \text{ if } w \notin E.$$

Notably, $\mathcal{E}_f$ is by definition coarser than $\mathcal{E}_\Phi$. So SL entails Self-Luminosity and therefore Self-Stability.

How should we think about SL? At first glance, the principle seems counter-intuitive, for reasons that we discussed in the end of Chapter 4. In particular, if the agent is sensitive to conditions beyond her own ken, then SL is in conflict with Weak Evidentialism, the principle that says it is always rationally permissible to plan to be less than fully confident in propositions that will not be entailed by one's evidence/knowledge. As discussed there, I think Weak Evidentialism is false. Sometimes it is a rational requirement to plan *ex ante* to be fully confident in propositions that you know will not be settled by your future evidence. It may be rationally permissible to deviate from that plan *ex post*, but we shouldn't forsake the best plan before we give it a shot.

There is however a different worry one may have about SL. One crucial assumption in Theorem 1 is that the set of available updating plans is closed under reflection. The thought is, if you can follow f , then you should be able to follow f^* just as well. Consider *Unmarked Clock*: the BCOND-plan tells you to rule out 9:48 if the time is 9:50, but to rule it in (with confidence 1/3) if the time is 9:49. Yet if you are already planning to adopt different credences depending on whether the time is 9:50 or 9:49, why not then follow the omniscient plan that tells you to be fully confident that the time is 9:50 when and only when it is 9:50 (and similarly for other times)?

The complaint, however, is that there is more to followability than sensitivity. In particular, whether a plan is followable may depend on not only its antecedents, but also its consequents, i.e. what it asks you to do under those antecedent conditions. Normal human beings can follow the plan that asks one to walk home if (one sees that) there are no onions in the grocery store; normal human beings can't follow the plan that asks one to change into an onion if there are no onions in the grocery store. Similarly, in *Unmarked Clock*, one might think that the BCOND-plan is in some sense more "available" than the omniscient plan (i.e. the META-COND-plan). The latter is closer to the plan of onion-transfiguration; it is asking you to do what is impossible to be achieved (given your limitations).

I have to confess that I don't share this intuition. It can't be that BCOND is the *only* updating plan that a rational agent can follow; any slight deviation from it is off the table. But once we allow other alternatives, it is unclear why updating plans that are (i) equally or less demanding in terms of the agent's sensitivity and (ii) better in accords with the truth, are somehow less followable. If anything, one might have thought that, perhaps for evolutionary reasons, we should be better at following *better* updating plans – plans that

would make our minds fit better with the world, and in turn bring the world more in line with our desires.

But I think this complaint is related to an important conceptual point, on which I will end. Recall the fundamental premise of cognitive decision theory: We should treat the problem of updating – or more precisely updating planning – as a cognitive decision problem. The rational updating plan is the optimal updating plan, where optimality is assessed on the basis of some decision-theoretic principles (e.g. dominance, utility maximization, etc.).

Now let me make two trivial observations about decision problems in general. First, what one should do in a decision problem depends on what options are available. I may want to have cereal for breakfast. But if I don't have cereal, then I *can't* be rationally required to have cereals – it is simply not an option.

Second, what options are available in a decision problem is not itself subject to rational assessment. When I find myself deliberating about what to have for breakfast, it is neither rational nor irrational that I don't have cereal. Maybe if I had thought about having cereal for breakfast the day before, then it would have been rational for me to make it the case that I had cereal now as an available option. But that's neither here nor there. More generally, we are interested in the question *What should one do in a decision problem*, not *What decision problem it is better for one to have*. The latter is a valid question, but not one that I am facing – I am deliberating about what to have for breakfast, and I've run out of cereal.

The same seems true of cognitive decision problems. From the decision-theoretic perspective, before you look at the clock, which updating plan you should adopt depends on which updating plans are available to you, and which updating plans are available to you is not itself subject to rational assessment (with respect to your problem). If your credences can be finely attuned to differences in your visual appearance, then good – you should plan to be confident that it is t if and only if it is t . But maybe you are not. Perhaps your credences are completely detached from how things appear to you. Then you should plan to retain your prior. It is unfortunate, but not irrational. People with good reflexes can catch things immediately when they fall; I can't. It would be great if my hands were in better sync with my eyes, though (I hope) the fact they are not is not a rational defect of mine.

But if this is right, then we have a problem. For what Theorem 1 shows is that, unless we make some substantive assumptions about what updating plans are available to an agent, in general the best updating plan recommended by cognitive decision theory is *not* going to be BCOND. If the set is closed under reflection, then the agent should update by conditionalizing on the true element of the partition to which she is most sensitive. If she is sensitive to what her evidence is and nothing else, then she should plan to update in accord with META-COND. If she is sensitive to the truths and nothing but the truths, then she should plan to be omniscient. If she can only follow updating plans that make one recommendation no matter what comes, then she is best served by planning to retain her prior. BCOND is optimal only for special agents and under

special conditions.

Perhaps this shows that we shouldn't have thought of updating plans as pills – or more generally as options that are evaluated on the basis of their outcomes. Perhaps one's credences at a time are only rationalizable by one's epistemic reasons at that time (e.g. evidence), and updating plans are rational insofar as they prescribe credences that will be rational at the time when they are adopted (cf. Closure). I am open to these possibilities. The main point of this and the previous chapter is that there is a tension between Bayesianism (in particular Conditionalization) on the one hand, and certain ways of thinking about updating plans on the other. Whether that means we should reject Bayesianism, or revise our conception of updating plans, is a question that I shall leave for future investigation.

Bibliography

- Ahmed, Arif, and Bernhard Salow. 2019. “Don’t look now”. *The British Journal for the Philosophy of Science* 70 (2): 327–350.
- Arnold, BC, and CA Roberston. 2019. “The Conditional Distribution of X Given $X = Y$ can be Almost Anything!” In *Advances on Theoretical and Methodological Aspects of Probability and Statistics*, 75–81. CRC Press.
- Arntzenius, Frank, Adam Elga, and John Hawthorne. 2004. “Bayesianism, Infinite Decisions, and Binding”. *Mind* 113 (450): 251–283. doi:10.1093/mind/113.450.251.
- Berker, Selim. 2008. “Luminosity Regained”. *Philosophers’ Imprint* 8:1–22.
- Berti, Patrizia, Enrique Miranda, and Pietro Rigo. 2017. “Basic ideas underlying conglomerability and disintegrability”. *International Journal of Approximate Reasoning* 88:387–400.
- Berti, Patrizia, and Pietro Rigo. 1999. “Sufficient conditions for the existence of disintegrations”. *Journal of Theoretical Probability* 12 (1): 75–86.
- Bertrand, Joseph. 1889. *Calcul des probabilités*. Gauthier-Villars.
- Billingsley, Patrick. 2008. *Probability and measure*. John Wiley & Sons.
- Blackwell, David, and Lester Dubins. 1962. “Merging of opinions with increasing information”. *The Annals of Mathematical Statistics* 33 (3): 882–886.
- Blackwell, David, and Lester E. Dubins. 1975. “On Existence and Non-Existence of Proper, Regular, Conditional Distributions”. *The Annals of Probability* 3 (5): 741–752.
- Blackwell, Kevin, and Daniel Drucker. 2019. “When propriety is improper”. *Philosophical Studies* 176 (2): 367–386.
- Briggs, R. A., and Richard Pettigrew. Forthcoming. “An Accuracy-Dominance Argument for Conditionalization”. *Nous*. doi:10.1111/nous.12258.

- Briggs, Rachael. 2009. “Distorted Reflection”. *Philosophical Review* 118 (1): 59–85. doi:10.1215/00318108-2008-029.
- Bronfman, Aaron. 2014. “Conditionalization and not knowing that one knows”. *Erkenntnis* 79 (4): 871–892.
- Brown, Peter M. 1976. “Conditionalization and Expected Utility”. *Philosophy of Science* 43 (3): 415–419. doi:10.1086/288696.
- Builes, David. 2020. “A Paradox of Evidential Equivalence”. *Mind* 129 (513): 113–127. doi:10.1093/mind/fzz046.
- . Forthcoming. “The Ineffability of Induction”. *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12753.
- Caie, Michael. 2013. “Rational Probabilistic Incoherence”. *Philosophical Review* 122 (4): 527–575. doi:10.1215/00318108-2315288.
- Chen, Robert. 1976. “Some finitely additive versions of the strong law of large numbers”. *Israel Journal of Mathematics* 24 (3): 244–259.
- Christensen, David. 1991. “Clever Bookies and Coherent Beliefs”. *Philosophical Review* 100 (2): 229–247. doi:10.2307/2185301.
- Cinlar, Erhan. 2011. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer New York. ISBN: 9780387878591. <https://books.google.com/books?id=7VtAx56r8CAC>.
- Das, Nilanjan. Forthcoming. “Externalism and Exploitability”. *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12742.
- Diaconis, Persi, and Sandy L. Zabell. 1982. “Updating Subjective Probability”. *Journal of the American Statistical Association* 77 (380): 822–830.
- Dorr, Cian, John Hawthorne, and Yoaav Isaacs. 2020. “Solving a Paradox of Evidential Equivalence”. *Mind*.
- Dorst, Kevin, et al. manuscript. “Deference Done Better”.
- Dubins, Lester E. 1977. “Measurable tail disintegrations of the Haar integral are purely finitely additive”. *Proceedings of the American Mathematical Society* 62 (1): 34–36.
- Dubins, Lester E, et al. 1975. “Finitely additive conditional probabilities, conglomerability and disintegrations”. *Annals of Probability* 3 (1): 89–99.
- Dubins, Lester E, and Karel Prikry. 1995. “On the existence of disintegrations”. *Séminaire de probabilités de Strasbourg* 29:248–259.

- Dubins, Lester E, et al. 2014. *How to gamble if you must: Inequalities for stochastic processes*. Courier Corporation.
- Earman, John. 1992. *Bayes or Bust?* Bradford.
- Easwaran, Kenny. 2019. “Conditional probabilities”.
- . 2013a. “Expected Accuracy Supports Conditionalization—and Conglomerability and Reflection”. *Philosophy of Science* 80 (1): 119–142. doi:10.1086/668879.
- . 2013b. “Expected accuracy supports conditionalization—and conglomerability and reflection”. *Philosophy of Science* 80 (1): 119–142.
- . 2013c. “Why countable additivity?” *Thought: A Journal of Philosophy* 2 (1): 53–61.
- Easwaran, Kenny Krishnan. 2008. “The foundations of conditional probability”. PhD thesis, University of California, Berkeley.
- Elga, Adam. 2013. “The Puzzle of the Unmarked Clock and the New Rational Reflection Principle”. *Philosophical Studies* 164 (1): 127–139. doi:10.1007/s11098-013-0091-0.
- Elga, Adam, and Agustin Rayo. Forthcoming. “Fragmentation and Logical Omniscience”. *Noûs*.
- Field, Hartry H. 1978. “Mental representation”. *Erkenntnis* 13 (1): 9–61.
- Finetti, Bruno de. 1970. *Theory of Probability*. New York: John Wiley.
- Fitelson, Branden, and James Hawthorne. 2010. “How Bayesian confirmation theory handles the paradox of the ravens”. In *The place of probability in science*, 247–275. Springer.
- Fraassen, Bas C. van. 1995. “Belief and the Problem of Ulysses and the Sirens”. *Philosophical Studies* 77 (1): 7–37. doi:10.1007/bf00996309.
- . 1989. *Laws and Symmetry*. Oxford University Press.
- Freiling, Chris. 1986. “Axioms of symmetry: throwing darts at the real number line”. *Journal of Symbolic Logic*: 190–200.
- Gaifman, Haim, and Anubav Vasudevan. 2012. “Deceptive Updating and Minimal Information Methods”. *Synthese* 187 (1): 147–178. doi:10.1007/s11229-011-0028-0.
- Gallow, J. Dmitri. 2019. “Diachronic Dutch Books and Evidential Import”. *Philosophy and Phenomenological Research* 99 (1): 49–80. doi:10.1111/phpr.12471.
- Goddard, L. 1977. “The paradoxes of confirmation and the nature of natural laws”. *The Philosophical Quarterly (1950-)* 27 (107): 97–113.

- Greaves, Hilary, and David Wallace. 2005. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility". *Mind* 115 (459): 607–632. doi:10.1093/mind/fzl607.
- . 2006. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility". *Mind* 115 (459): 607–632. doi:10.1093/mind/fzl607.
- Hamkins, Joel David, et al. 2015. "Is the dream solution of the continuum hypothesis attainable?" *Notre Dame Journal of Formal Logic* 56 (1): 135–145.
- Hedden, Brian. 2013. "Incoherence Without Exploitability". *Noûs* 47 (3): 482–495. doi:10.1111/j.1468-0068.2011.00842.x.
- Hempel, Carl G. 1945. "Studies in the Logic of Confirmation (I.)" *Mind* 54 (213): 1–26.
- Hill, Bruce M., and David Lane. 1985. "Conglomerability and Countable Additivity". *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 47 (3): 366–379. ISSN: 0581572X. <http://www.jstor.org/stable/25050554>.
- Hoek, Daniel. 2020a. "Chance and the Continuum Hypothesis". *Philosophy and Phenomenological Research*.
- . 2020b. "Minimal rationality and the web of questions".
- Holguin, Ben. 2020. "Thinking, guessing, and believing". *Manuscript, NYU*.
- Howson, C., and P. Urbach. 2006. *Scientific Reasoning: The Bayesian Approach*. Philosophy Series. Open Court. ISBN: 9780812695786. <https://books.google.com/books?id=riQfAQAAIAAJ>.
- Howson, Colin. 2014. "Finite additivity, another lottery paradox and conditionalisation". *Synthese* 191 (5): 989–1012.
- Hughes, R. I. G., and Bas C. van Fraassen. 1984. "Symmetry Arguments in Probability Kinematics". *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1984:851–869. ISSN: 02708647. <http://www.jstor.org/stable/192543>.
- Huttegger, Simon M. 2013. "In defense of reflection". *Philosophy of Science* 80 (3): 413–433.
- Huttegger, Simon M. 2017. *The Probabilistic Foundations of Rational Learning*. Cambridge University Press.
- Huttegger, Simon M, and Michael Nielsen. 2020. "Generalized learning and conditional expectation". *Philosophy of Science* 87 (5): 868–883.
- Jaynes, Edwin T. 2003. *Probability theory: The logic of science*. Cambridge university press.
- Joyce, James M. 2009. "Accuracy and coherence: Prospects for an alethic epistemology of partial belief". In *Degrees of belief*, 263–297. Springer.
- . 1999. *The foundations of causal decision theory*. Cambridge University Press.

- Juhl, Cory, and Kevin T Kelly. 1994. “Realism, convergence, and additivity”. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1994:181–189. 1. Philosophy of Science Association.
- Kadane, Joseph B., Mark J. Schervish, and Teddy Seidenfeld. 1996. “Reasoning to a Foregone Conclusion”. *Journal of the American Statistical Association* 91 (435): 1228–1235. ISSN: 01621459.
- Khan, M Ali, et al. 2020. “Modeling large societies: Why countable additivity is necessary”. *Journal of Economic Theory* 189:105102.
- Kolmogorov, A.N., N. Morrison, and A.T. Bharucha-Reid. 1956. *Foundations of the Theory of Probability*. AMS Chelsea Publishing Series. Chelsea Publishing Company. <https://books.google.com/books?id=EmetAAAAMAAJ>.
- Kolodny, Niko. 2007. “IX—How Does Coherence Matter?” In *Proceedings of the Aristotelian Society (Hardback)*, 107:229–263. 1pt3. Wiley Online Library.
- . 2008. “Why be disposed to be coherent?” *Ethics* 118 (3): 437–463.
- Lasonen-Aarnio, Maria. 2014. “Higher-order evidence and the limits of defeat”. *Philosophy and Phenomenological Research* 88 (2): 314–345.
- . 2015. “New rational reflection and internalism about rationality”. *Oxford studies in epistemology* 5:145–171.
- Lewis, David. 1971. “Immodest inductive methods”. *Philosophy of Science* 38 (1): 54–63.
- Lewis, David K. 1980. “A Subjectivist’s Guide to Objective Chance”. In *Studies in Inductive Logic and Probability, Volume II*, ed. by Richard C. Jeffrey, 263–293. Berkeley: University of California Press.
- . 1999. “Why Conditionalize”. In *Philosophy of Probability: Contemporary Readings*, ed. by Antony Eagle, 403–407. Routledge.
- Lin, Hanti, and Kevin T Kelly. 2012. “Propositional reasoning that tracks probabilistic reasoning”. *Journal of philosophical logic* 41 (6): 957–981.
- Maddy, Penelope. 2017. “Set-theoretic foundations”. *Foundations of mathematics* 690:289.
- Meehan, Alexander, and Snow Zhang. manuscript. “Instrumentalism and Evidentialism in Updating”.
- . Forthcoming. “Kolmogorov Conditionalizers Can Be Dutch Booked”. *Review of Symbolic Logic*: 1–36. doi:10.1017/s1755020320000295.
- . 2020. “The Borel-Kolmogorov paradox is your paradox too: A puzzle for conditional physical probability”. *Philosophy of Science*.

- Moss, Sarah. 2015. "Time-slice epistemology and action under indeterminacy". *Oxford studies in epistemology* 5 (5): 172–193.
- Myrvold, Wayne C. 2015. "You Can'T Always Get What You Want: Some Considerations Regarding Conditional Probabilities". *Erkenntnis* 80 (3): 573–603. doi:10.1007/s10670-014-9656-3.
- Nielsen, Michael. Forthcoming. "Accuracy-Dominance and Conditionalization". *Philosophical Studies*: 1–20. doi:10.1007/s11098-020-01598-6.
- Oxtoby, J.C. 1971. *Measure and Category: A Survey of the Analogies between Topological and Measure Spaces*. Graduate Texts in Mathematics. Springer New York. ISBN: 9780387900254. <https://books.google.com/books?id=-TjvAAAAMAAJ>.
- Paul, Sarah K. 2014. "Diachronic incontinence is a problem in moral philosophy". *Inquiry* 57 (3): 337–355.
- Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- Pettigrew, Richard. 2016a. *Accuracy and the Laws of Credence*. Oxford University Press.
- . 2016b. *Accuracy and the Laws of Credence*. Oxford University Press UK.
- . 2020. *Dutch Book Arguments*. Cambridge University Press.
- Pomatto, Luciano, and Alvaro Sandroni. 2018. "An axiomatic theory of inductive inference". *Philosophy of Science* 85 (2): 293–315.
- Purves, Roger A, and William D Sudderth. 1983. "Finitely additive zero-one laws". *Sankhyā: The Indian Journal of Statistics, Series A*: 32–37.
- . 1976. "Some finitely additive probability". *The Annals of Probability*: 259–276.
- Ramsey, F. 1926. "Truth and Probability". In *Philosophy of Probability: Contemporary Readings*, ed. by Antony Eagle, 52–94. Routledge.
- Rescorla, Michael. 2018. "A Dutch Book Theorem and Converse Dutch Book Theorem for Kolmogorov Conditionalization". *Review of Symbolic Logic* 11 (4): 705–735. doi:10.1017/s1755020317000296.
- . Forthcoming(a). "An Improved Dutch Book Theorem for Conditionalization". *Erkenntnis*: 1–29. doi:10.1007/s10670-020-00228-1.
- . Forthcoming(b). "On the Proper Formulation of Conditionalization". *Synthese*: 1–31. doi:10.1007/s11229-019-02179-9.
- Rinard, Susanna. 2014. "A new bayesian solution to the paradox of the ravens". *Philosophy of Science* 81 (1): 81–100.

- Salow, Bernhard. 2018. "The Externalist's Guide to Fishing for Compliments". *Mind* 127 (507): 691–728. doi:10.1093/mind/fzw029.
- Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley Publications in Statistics.
- Schafer, Karl. 2014. "Doxastic Planning and Epistemic Internalism". *Synthese* 191 (12): 2571–2591. doi:10.1007/s11229-014-0412-7.
- . 2018. "Epistemic planning, epistemic internalism, and luminosity". *Metaepistemology*: 179.
- Schervish, Mark J, Teddy Seidenfeld, and Joseph B Kadane. 1984. "The extent of non-conglomerability of finitely additive probabilities". *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 66 (2): 205–226.
- Schoenfield, Miriam. 2018. "An Accuracy Based Approach to Higher Order Evidence". *Philosophy and Phenomenological Research* 96 (3): 690–715. doi:10.1111/phpr.12329.
- . 2017. "Conditionalization Does Not Maximize Expected Accuracy". *Mind* 126 (504): 1155–1187. doi:10.1093/mind/fzw027.
- Seidenfeld, T., Mark J. Schervish, and Joseph B. Kadane. 2014. "Non-Conglomerability for Countably Additive Measures That Are Not κ -Additive". *Review of Symbolic Logic* 10 (2): 284–300. doi:10.1017/s1755020316000344.
- Seidenfeld, Teddy, Mark J. Schervish, and Joseph B. Kadane. 2001. "Improper Regular Conditional Distributions". *Annals of Probability* 29, no. 4 (): 1612–1624. doi:10.1214/aop/1015345764. <https://doi.org/10.1214/aop/1015345764>.
- Seidenfeld, Teddy, Mark J Schervish, and Joseph B Kadane. 1998. "Non-conglomerability for finite-valued, finitely additive probability". *Sankhyā: The Indian Journal of Statistics, Series A*: 476–491.
- Shafer, Glenn. 1985. "Conditional probability". *International Statistical Review/Revue Internationale de Statistique*: 261–275.
- Shafer, Glenn, and Vladimir Vovk. 2006. "The Sources of Kolmogorov's" Grundbegriffe"". *Statistical Science*: 70–98.
- Singpurwalla, Nozer D, and Andrew Swift. 2001. "Network reliability and Borel's paradox". *The American Statistician* 55 (3): 213–218.
- Srinivasan, Amia. 2020. "Radical externalism". *Philosophical Review* 129 (3): 395–431.
- Staffel, Julia. 2019. *Unsettled Thoughts: A Theory of Degrees of Rationality*. Oxford University Press.

- Stalnaker, Robert. 2015. “Luminosity and the KK Thesis”. In *Externalism, Self-Knowledge, and Skepticism: New Essays*, ed. by Sanford Goldberg, 19–40. Cambridge University Press.
- . 1991. “The problem of logical omniscience, I”. *Synthese* 425–440.
- Titelbaum, Michael. 2015. “Fundamentals of Bayesian epistemology”. *Unpublished manuscript*.
- Weisberg, Jonathan. 2015a. “Updating, undermining, and independence”. *The British Journal for the Philosophy of Science* 66 (1): 121–159.
- . 2015b. “You’ve come a long way, Bayesians”. *Journal of Philosophical Logic* 44 (6): 817–834.
- Williamson, Jon. 1999. “Countable additivity and subjective probability”. *The British Journal for the Philosophy of Science* 50 (3): 401–416.
- Williamson, Timothy. 2011. “Improbable knowing”. *Evidentialism and its Discontents* 147:164.
- . 2000. *Knowledge and its Limits*. Oxford University Press.
- . 2014. “Very improbable knowing”. *Erkenntnis* 79 (5): 971–999.
- Worsnip, Alex. 2018. “The Conflict of Evidence and Coherence”. *Philosophy and Phenomenological Research* 96 (1): 3–44. doi:10.1111/phpr.12246.
- Xue, Chao, et al. 2020. “Precision measurement of the Newtonian gravitational constant”. *National Science Review* 7 (12): 1803–1817.
- Yablo, Stephen. 2014. *Aboutness*. Princeton University Press.
- Yalcin, Seth. 2016. “Belief as question-sensitive”. *Philosophy and Phenomenological Research* 97 (1): 23–47.
- . Forthcoming. “Fragmented But Rational”. In *The Fragmented Mind*. Oxford University Press.
- . 2010. “Probability operators”. *Philosophy Compass* 5 (11): 916–937.
- Zame, Alan. 1988. “Conglomerability and finite partitions”. *Proceedings of the American Mathematical Society* 102 (1): 165–168.

Appendices

Appendix A

Chapter 1

1 Global Coherence

Fact 1. Fix a coherent credal state C . C is globally coherent if and only if there exists a coherent conditional credence function $\tilde{C} : \mathcal{F}_C \times \mathcal{F}_C \setminus \{\emptyset\} \rightarrow \mathbb{R}$ such that for all $A \in \mathcal{F}_C$, partition $\mathcal{E} \subset \mathcal{F}_C$ and non-empty $B \in \mathcal{E}$,

$$\tilde{C}(A|B) = C_{\mathcal{E}}(A|B)$$

Proof. The $\Leftarrow$ -direction is straightforward: Suppose there is a coherent conditional credence function $\tilde{C}$ that satisfies the above condition. Let $\mathcal{E}, \mathcal{E}' \subset \mathcal{F}_C$ and $E \in \mathcal{E}, E' \in \mathcal{E}'$ with $E' \subset E$ and $E' \neq \emptyset$. Let $A \in \mathcal{F}_C$. Then

$$C_{\mathcal{E}}(A \cap E' | E) = \tilde{C}(A \cap E' | E) = \tilde{C}(A | E' \cap E) \tilde{C}(E' | E) = C_{\mathcal{E}'}(A | E') C_{\mathcal{E}}(E' | E).$$

So C is globally coherent.

For the converse: Suppose C is globally coherent. For each non-empty $B \in \mathcal{F}_C$, let $\mathcal{E}_B = \{B, B^c\}$. Consider a conditional credence function $\tilde{C} : \mathcal{F}_C \times \mathcal{F}_C \setminus \{\emptyset\} \rightarrow \mathbb{R}$ given by

$$\tilde{C}(A|B) = C_{\mathcal{E}_B}(A|B).$$

Since C is coherent, for every $B \in \mathcal{F}_C \setminus \{\emptyset\}$, $C_{\mathcal{E}_B}(\cdot|B)$ is a finitely additive probability function with $C_{\mathcal{E}_B}(B|B) = 1$. So it remains to check that (i) $\tilde{C}$ satisfies the Multiplication Axiom and (ii) $\tilde{C}(\cdot|B)$ agrees with $C_{\mathcal{E}}(\cdot|B)$ for any arbitrary $\mathcal{E}$ that contains B . For (i): Let $F, G, H \in \mathcal{F}_C$ with $G \cap H \neq \emptyset$. Let

$\mathcal{E} = \{H, H^c\}$ and $\mathcal{E}' = \{G \cap H, (G \cap H)^c\}$. Then

$$\tilde{C}(F \cap G|H) = \tilde{C}(F \cap (G \cap H)|H) = C_{\mathcal{E}}(F \cap (G \cap H)|H)$$

and

$$\tilde{C}(F|G \cap H)\tilde{C}(G|H) = C_{\mathcal{E}'}(F|G \cap H)C_{\mathcal{E}}(G \cap H|H)$$

By Global Coherence,

$$C_{\mathcal{E}}(F \cap (G \cap H)|H) = C_{\mathcal{E}'}(F|G \cap H)C_{\mathcal{E}}(G \cap H|H).$$

So $\tilde{C}(F \cap G|H) = \tilde{C}(F|G \cap H)\tilde{C}(G|H)$. So $\tilde{C}$ satisfies the Multiplicative Axiom.

Lastly, to verify (ii): Let $A, B \in \mathcal{F}_C$ with $B \neq \emptyset$. Let $\mathcal{E}$ be a partition that contains $B \in \mathcal{E}$. Note that $B \in \mathcal{E} \cap \mathcal{E}_B$. As noted in Chapter 1, Global Coherence entails that

$$C_{\mathcal{E}}(\cdot|B) = C_{\mathcal{E}_B}(\cdot|B).$$

So $\tilde{C}(\cdot|B) = C_{\mathcal{E}_B}(\cdot|B) = C_{\mathcal{E}}(\cdot|B)$, as desired. □

2 Freiling and Lévy

Freiling (1986) proved that $\neg\text{CH}$ is equivalent to the following principle:

Freiling's Symmetry Axiom. $\forall f : \mathbb{R} \rightarrow \mathbb{R}_{\aleph_0}, \exists x \exists y (y \notin f(x) \wedge x \notin f(y))$

where $\mathbb{R}_{\aleph_0}$ denote the set of all countable subsets of $\mathbb{R}$.¹

Freiling argued for FSA on the basis of the following thought experiment:

Double Darts. Two infinitesimal darts are thrown onto the real line at random. The first dart hits X and the second dart hits Y . Suppose $X \neq Y$ and the two dart throws are probabilistically independent of each other.

The argument then proceeds by a *reductio*. Suppose CH is true. Under ZFC, this means that there is a well-ordering $<$ on the reals such that the set of predecessors of every real is countable. It follows that, for

¹To briefly rehearse his proof: Note that FSA fails just in case there is a mapping from real numbers to countable subsets of real numbers such that, for every pair of real numbers, at least one must be in the image of the other. To see $\text{FSA} \Rightarrow \neg\text{CH}$, suppose CH is true. Let f be the map $x \mapsto \{y : y < x\}$, where $<$ is a well-ordering on reals, which is possible given ZFC. Given CH, for every x , $f(x)$ is countable. By the definition of well-orders, for every pair of real numbers, either $x \leq y$ or $y \leq x$. So for every x, y , either $y \in f(x)$ or $x \in f(y)$, *contra* FSA.

For the other direction: Suppose CH is false. Then $\mathbb{R}$ has a proper subset A of cardinality $\aleph_1$. Let f be any mapping from reals to countable subsets of reals. Then the image of A under f would also have cardinality $\aleph_1$. It follows that there is some y not in the image of A (in particular $y \notin f(x)$ for all $x \in A$), and since $f(y)$ is countable, there is $x \in A$ such that $x \notin f(y)$ either. So FSA is true.

any value of X , Y is not a predecessor of X with probability 1 (since the Lebesgue measure of a countable set is 0). Similarly X is not a predecessor of Y with probability 1 for any value of Y . So it would appear that, before the darts are thrown, one can predict with certainty that neither will hit a number that is a predecessor of the other. Yet this is absurd, since it follows from the definition of a well-ordering that one of them must be the predecessor of the other.

Notably, the last step relies on a version of the conglomerability principle. In fact Freiling admitted to this explicitly :

To reject our axioms [FSA], it would seem also necessary to reject one of the following four self-evident philosophical principles.

1. Choosing reals at random is a physical reality, or at least an intuition mathematics should embrace to the extent possible.
2. A fixed [countable] set predictably will not be hit by a random dart.²
3. If an accurate Yes-No prediction can always be made after a preliminary event takes place (e.g., the first dart is thrown) and, no matter what the outcome of that event, the prediction is always the same, then the prediction is also in some sense accurate before the preliminary event.
4. The real number line cannot tell the order of the darts. (Freiling 1986, p.199)

Freiling continued to remark:

Principle 3 is not meant to be a mathematical statement of the Lebesgue measurability of a certain type of set. Rather, it is an expression of an obvious, almost physical intuition concerning the inherently nonmathematical notions of prediction, accuracy, and time independence.

As noted in Chapter 1, there is a close parallel between Freiling's *Double Darts* example, and the following case attributed to Paul Lévy by Cantelli (Seidenfeld et al. 1998), which is often taken to have convincingly demonstrated the irrationality of uniform measures over $\mathbb{N} \times \mathbb{N}$ (I have adapted the case to make the parallel more explicit):

*Double Darts**. Two darts are thrown onto the $\mathbb{N} \times \mathbb{N}$ grid. The first dart hits a natural number X , and the second dart hits a natural number Y . Suppose $X \neq Y$ and the two dart throws are probabilistically independent of each other.

²In his original formulation, Freiling used “a fixed Lebesgue measure zero set” instead of “a fixed countable set.” However, since every countable set has measure zero, this principle is weaker and sufficient for the argument to go through.

Now suppose P is a uniform probability distribution over $\mathbb{N} \times \mathbb{N}$. In particular, $P(Y = n|X = m) = 0 = P(X = m|Y = n)$, for all $m, n \in \mathbb{N}$. Then for any x, y , one has $P(Y \leq X|X = x) = 0$ and $P(X \geq Y|Y = y) = 0$. Assuming conglomerability, it follows that $P(Y \leq X \vee X \leq Y) = 0$, which contradicts the normalization assumption that $P(\Omega) = 1$.

Lévy’s example shows that the following four claims form an inconsistent quartet:

1. Choosing *natural numbers* at random is a physical reality, or at least an intuition mathematics should embrace to the extent possible.
2. A fixed *finite* set predictably will not be hit by a random dart.
3. If an accurate Yes-No prediction can always be made after a preliminary event takes place (e.g., the first dart is thrown) and, no matter what the outcome of that event, the prediction is always the same, then the prediction is also in some sense accurate before the preliminary event.
4. The *natural number* line cannot tell the order of the darts.

Lévy’s inconsistent quartet appears exactly analogous to Freiling’s. The only difference is that “the reals” are substituted with “the natural numbers,” and “countable” with “finite.” It is thereby intriguing that the CA-theorist responds to Lévy’s quartet by rejecting 1, and to Freiling’s quartet by rejecting 3 (in particular they reject 3 when the events correspond to *non-measurable* sets (cf. Hamkins et al. (2015))).

3 Builes’ paradox

It may be worth mentioning a different yet related paradox, introduced by Builes (2020). Suppose a countable infinity of physically isomorphic fair coins are tossed. Name one of them “Chloe.” What is the probability that Chloe will land heads, given that (i) only finitely many coins will land heads; (ii) finitely many coins *except Chloe* will land heads? Intuitively, the answer to (i) seems to be 0, whereas the answer to (ii) seems to be 1/2. But (i) and (ii) are the same conditions (up to necessary equivalence). So one must give.

A tempting response to Builes’ paradox is that the 0-intuition depends on some form of conglomerability-based reasoning (cf. Dorr et al. (2020)). The idea is that, we have the 0-intuition because the following is true: For every n , given that exactly n many coins will land heads, the probability that Chloe will land heads is 0. More precisely, let F denote the proposition that finitely many coins will land heads, F_n the proposition that exactly n coins will land heads and H_C the proposition that Chloe will land heads. Let P be a rational prior. Then

Local Zero. For all n , $P(H_C|F_n) = 0$.

We are tempted to reason from Local Zero to the conclusion that $P(H_C|F) = 0$ because F is simply the disjunction of F_n (which are mutually exclusive), and we are tempted to think that rational credence should be conglomerable – if $P(H_C|F_n) = 0$ for all n , then $P(H_C|F) = 0$. But theorists (in particular anti-CA theorists) who have learnt to live with inconglomerability would of course reject this last step of reasoning. In *Double Darts**, for all n , $P(X > Y|Y = n) = 1$ and $P(X < Y|X = n) = 1$ but it does not follow that $P(X > Y) = 1$ and $P(Y > X) = 1$. Similarly, in Builes’ paradox, even though $P(H_C|F_n) = 0$ for every n , it does not follow that $P(H_C|F) = 0$.

The above diagnosis is an error theory. It says that our 0-intuition is mistaken, and offers an explanation why we have the intuition nonetheless: The answer $P(H|F) = 0$ seems plausible because the principle of conglomerability seems plausible, until subject to further philosophical scrutiny.

My main reservation about this error theory is that the explanation does not seem to generalize to all versions of the paradox. For instance, suppose ω_1 -many coins are flipped, and one is told (i) *exactly* countably infinitely many coins will land heads; (ii) exactly countably infinitely many coins other than Chloe will land heads. Again, we seem to have the intuition that conditional on (i), the rational probability that Chloe will land heads is 0, whereas conditional on (ii), the rational probability that Chloe will land heads is $1/2$. According to the error theory, the 0-intuition is mistaken and we had that intuition because we were appealing to some instance of the conglomerability principle. But it is not immediately clear what is the relevant partition with respect to which the conglomerability principle is invoked in this case.

Instead, I think Builes’ paradox is connected to Lévy’s and Freiling’s paradoxes in a slightly different way. To see how, let’s start with a slightly different problem:

Double Darts and Coin. Two darts are thrown onto the $\mathbb{N} \times \mathbb{N}$ grid. The first dart hits a natural number X ; the second dart hits a natural number Y . Suppose that for all n ,

- $P(X = n) = P(Y = n) = 0$;
- $P(X > Y|Y = n) = 1, P(Y > X|X = n) = 1$.

Someone will decide what to do with a fair coin based on the values of X and Y . If $X > Y$, then the coin is tossed; if $X < Y$, then the coin is placed tails up; if $X = Y$, then the coin is destroyed. Let H be the proposition that the coin is tossed and lands heads. Given the information of the set-up, what is the rational probability of H ?³

Let’s set aside the answer to this question for now (the reader may want to pause here to check their intuitions.) The crucial point is that the constraints we imposed on P entail the following:

³While I speak of “the rational probability of H ,” I am open to there being more than one probability an ideally rational agent may assign to H , given the information specified above and nothing else.

1. $\forall n, P(H|Y = n) = P(H|X > Y, Y = n) = 1/2$;
2. $\forall n, P(H|X = n) = P(H|Y > X, X = n) = 0$.

So an ideally rational agent with credence function P should be $1/2$ confident in H if she learns the position of the second dart and nothing else, but certain in $\neg H$ if she learns the position of the first dart and nothing else.

Now let's go back to Builes' paradox. Suppose the coins are assigned labels $1, 2, \dots$ at random. *A priori*, one does not know the number of Chloe. Let N be the number of the last coin that will land heads. The assumption that only finitely many coins land heads is equivalent to the assumption that N is finite. Let C be the number of Chloe. Pre-theoretically, one might think: For all n , conditional on $N = n$, it is rational to be certain that $C > N$, whereas conditional on $C = n$, it is rational to be certain that $N > C$. That is,

3. $\forall n, P_F(C > N|N = n) = 1, P_F(N > C|C = n) = 1$.

Now, by definition, $C > N$ only if Chloe lands tails, i.e. H_C is false. On the other hand, for any n , $N > C \& C = n$ appears probabilistically independent of H_C ; by assumption, the tosses are independent of each other, so learning that Chloe is labelled n and some coin of a larger label will land heads should not affect the probability of whether Chloe (i.e. the n -th coin) will land heads or tails. In other words, the following combination seems rationally permissible:

4. $\forall n, P_F(H_C|C > N, N = n) = 0, P_F(H_C|C < N, C = n) = 1/2$.

Putting together 1 and 2, we have

5. $\forall n, P_F(H_C|N = n) = P_F(H_C|C > N, N = n)P_F(C > N|N = n) = 0$;
6. $\forall n, P_F(H_C|C = n) = P_F(H_C|N > C, C = n)P_F(N > C|C = n) = 1/2$,

which are exactly analogous to 1 and 2. To this extent, the problem appears formally equivalent to *Double-Darts-Coins*. In both cases, there are two partitions $\mathcal{E}$ and $\mathcal{E}'$ of some set of logical possibilities and some proposition A , such that, for some rational P with $P(\bigcup \mathcal{E}) = P(\bigcup \mathcal{E}') = 1$, it appears rationally permissible that

$$\forall E \in \mathcal{E}, P(A|E) = 1/2, \text{ and } \forall E' \in \mathcal{E}', P(A|E') = 0.$$

So, on my diagnosis, Builes' paradox turns on *two* applications of the conglomerability principle that lead to incompatible answers; one says that $P_F(H_C)$ should be $1/2$, the other says that it should be 0.

(One advantage of my diagnosis of the paradox is that it generalizes to cases where more than countably infinitely many coins are tossed. For instance, suppose ω_1 -many fair coins are tossed and the information

learnt is that exactly countably infinitely many lands heads. Let F_ω denote this proposition. Then intuitively

$$\forall \alpha < \omega_1, P_F(C > N | N = \alpha) = 1, P_F(N > C | C = \alpha) = 1$$

which entails that

$$\forall \alpha < \omega_1, P_F(H_C | N = \alpha) = 0, P_F(H_C | C = \alpha) = 1/2.$$

Before concluding this section, let me address one objection.

Objection. Suppose the label of Chloe is given, say n . In this case, we still have the intuition that, given that only finitely many coins land heads, the probability that Chloe (i.e. the n -th coin) lands heads is 0. But my diagnosis says that $P_F(H_C | C = n) = 1/2$. So my explanation is incomplete; it does not explain why the 0-intuition persists despite knowledge of the value of C .

Reply. I think we have the intuition that $P_F(H_C | C = n) = 0$ because we intuitively think that whether Chloe lands heads is probabilistically independent of its label, and we reason that $P_F(H_C) = 0$ by applying the principle of conglomerability to the partition $\{N = n\}_{n \in \mathbb{N}}$, i.e.

$$P_F(H_C | C = n) = P_F(H_C) = P_F(H_C | N = n) = 0.$$

It is akin to reasoning in *Double Darts** that $P(X > Y | Y = n) = 0$ on the basis that $P(X > Y | X = n) = 0$ for all n . But once one has internalized the paradoxical phenomenon of inconglomerability, one would realize that both steps are mistaken: $P_F(H_C | C = n)$ may not equal $P_F(H_C | N = n)$, and neither needs to be equal to $P_F(H_C)$.

Appendix B

Chapter 2

In this Appendix, I briefly discuss how Fact 1 of Chapter 2 relates to a recent result in Builes (forthcoming). For simplicity, let's fix the sample space Ω to be the set of infinite binary strings, i.e. $\Omega = \{0, 1\}^{\mathbb{N}}$. We can think of each world $w = (w_1, w_2, \dots) \in \Omega$ as representing the outcome of an infinite sequence of iid coin tosses; $w_i = 1$ if the coin lands heads on the i -th toss, and 0 if the coin lands tails on the i -th toss. We say $w \sim_n v$ if w and v agree on the first n tosses. Note that $\sim_n$ is an equivalence relation for all $n \geq 1$. Let $w^n = \{v \in \Omega : w \sim_n v\}$. Let C a prior conditional credence function over Ω with $\mathcal{F}_C$ being the sigma-algebra generated by sets of the form w^n . A straightforward adaptation of the proof of Fact 1 shows that, if C is weakly regular, then there is $E \in \mathcal{F}_C$ such that $C(\cdot|E)$ is only finitely but not countably additive. In other words, Weak Regularity is incompatible with CA-COND, the principle that $C(\cdot|E)$ is countably additive for every $E \in \mathcal{F}_C$.

Builes (forthcoming) proves that Weak Regularity is incompatible with another principle. Say C is **skeptical about** w if $C(w|w^n) = 0$ for all n ; C is **skeptical** if it is skeptical about every $w \in \Omega$. Say C is **non-skeptical** if it is not skeptical, i.e. there is $w \in \Omega$ and n such that $C(w|w^n) > 0$ (since $w^{n+k} \subset w^n$ and $w \in w^{n+k}$, we have $C(w|w^{n+k}) > 0$ if $C(w|w^n) > 0$). Let *Non-Skepticism* be the following principle

Non-Skepticism. It is a rational requirement that C be non-skeptical.

Intuitively, skepticism about w means that one is disposed to be confident that w is *false*, despite getting (increasingly) strongly confirming evidence that w is true. This seems *prima facie* irrational. If you see the coin landing heads a trillion times, then, intuitively, you should give *some* credence to the possibility that it will always land heads.

Fact 2 (Builes, forthcoming). *If C is weakly regular, then C is skeptical.*

What shall we make of Builes' result? For what it's worth, I am inclined to think that it is rationally permissible to have $C(w|w^n) = 0$ for some w and n , i.e. I am a Skeptic in Builes' sense. Builes gave the following argument in favor of Non-Skepticism (in our setting, we can think of laws as individual worlds w and evidence as w^n),

If the experimental evidence E for [the law] L was so good that every physicist unanimously endorsed L , then at the very least we shouldn't be certain that L is false! Giving up on [Non-Skepticism] corresponds to a kind of skepticism about inductive reasoning, since inductive reasoning is supposed to align with how scientists actually reason. In fact, denying [Non-Skepticism] is analogous to claiming that induction is no better than counter-induction. After all, if [Non-Skepticism] is false, having the exact opposite views as the scientific community (with respect to the laws of physics) is perfectly rational.

I have two reservations about this argument. First, Non-Skepticism strikes me as too strong. Even if the coin has landed heads a trillion times, there are still infinitely many tosses to come. The odds of observing at least one tails amongst those infinitely many tosses is infinitely higher than the odds of observing no tails at all. Thus, it seems to me that one can be rationally certain that the coin won't land heads, even after observing a trillion heads in a row.

More generally, Builes motivated Non-Skepticism on the basis of scientific practice. The thought seems to be: there may be some strong evidence that would make every physicist fully confident w is true; in that hypothetical scenario, one should not be fully confident that w is false. But one might think that physicists can never be *rationally* certain that w is true, on the basis of a piece of evidence that leaves open whether the actual world is w or some other possibility in an uncountable set. The latest recommended value of the gravitational constant by the Committee on Data for Science and Technology is $(6.67408 \pm 0.00031) \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ (Xue et al. 2020). This is a very small range, but it still massively underdetermines whether the gravitational constant is *exactly* $6.67408 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ as opposed to $6.67409 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$. Is one rationally required to give some positive credence to the hypothesis that the gravitational constant is some precise number r within that range? The answer is not obvious to me.

I think the above consideration suggests a weaker principle than Non-Skepticism. Say C is **weakly non-skeptical about** w just in case

$$C(w^{n+k}|w^n) > 0, \text{ for any } k > 0.$$

Again it is straightforward to verify that $C(w^{n+k}|w^n) > 0$ only if $C(w^{n+k}|w^m) > 0$ for all $m > n$.) Intuitively, weak non-skepticism about w means that, given compelling evidence compatible with w , one does not assign zero probability to any set of hypotheses that approximate w up to some finite degree of precision. For instance, suppose the actual world is one where the coin always lands heads. Then after observing a trillion heads, a weakly non-skeptical agent won't rule out the possibility that the coin will land heads on another trillion tosses, or another trillion of trillion tosses, etc. Let Weak Non-Skepticism be the principle that it is a rational requirement that C be weakly non-skeptical. It appears to me that Weak Non-Skepticism is sufficient for vindicating the rationality of scientific practice.

It is straightforward to check that, even though Weak Regularity is incompatible with Non-Skepticism, it is compatible with Weak Non-Skepticism. In particular, let C be the fair coin measure with $C(w|w \vee v) = 0.5$ for all w and v . Then for any string w and $n, k \in \mathbb{N}$, we have

$$C(w^{n+k}|w^n) = \frac{C(w^n)2^{-k}}{C(w^n)} = 2^{-k} > 0.$$

So C is weakly regular and weakly non-skeptical.

Second, Builes claims that denying Non-Skepticism “is analogous to claiming that induction is no better than counter-induction.” Again I find this claim too strong. One can distinguish between two ways of showing that induction is better than counter-induction: (i) it increases our likelihood of believing the truths; (ii) it increases our likelihood of disbelieving the falsehoods. If C is skeptical, then C is inaccurate in the sense that it assigns full credence to falsehoods, despite increasing evidence. But there is a different sense in which C is doing better than its counter-inductive counterpart. To make this idea precise: say C is **weakly inductive about** w if, when w is not the actual world, C thinks it will be able to eventually confirm that w is *false* with certainty, i.e.

$$C\{v : v \neq w \Rightarrow \lim_{n \rightarrow \infty} C(w|v^n) = 0\} = 1.$$

Let *Weak Inductivism* be the view that C should be weakly inductive about every w . One can show that, even though Weak Regularity is incompatible with Non-Skepticism, it is compatible with Weak Inductivism. In particular, let C be countably additive and weakly regular. Then $C\{v : v \neq w\} = 1$ and by Lévy's 0-1 law, for any hypothesis $H \in \mathcal{F}_C$,

$$C\{v : \lim_{n \rightarrow \infty} C(H|v^n) = 1_H(v)\} = 1.$$

By letting H be the singleton hypothesis $\{w\}$, we arrive at the desired result that C is weakly inductive. So

a Skeptic does not have to believe that induction is as good as counter-induction – induction at least allows one to be confident that one will come to know what is false, even though it may not help one figure out what is (exactly) true.

To take stock: although Weak Regularity rules out Non-Skepticism, it does not rule out Weak Non-Skepticism or Weak Inductivism, which one might think is all that it requires to vindicate the rationality of induction or our best scientific practice.

Now let's go back to CA. Recall that Weak Regularity entails the negations of both CA-COND and Non-Skepticism. A natural (and to my knowledge still open) question is whether there is any logical relation between CA-COND and Non-Skepticism. There are some results in the vicinity (cf. Pomatto and Sandroni (2018)). To state these results, however, we need to introduce two more technical notions.

1. *Inductivism*. Before we defined weak inductivism: C is weakly inductive about w just in case it assigns probability one to the proposition that, in the limit, it will be able to rule out worlds that are non-actual. Say C is **inductive about** w if it will become increasingly confident in w given increasingly confirming evidence, i.e.

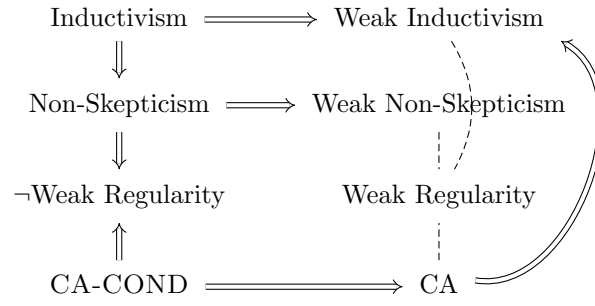
$$\lim_{n \rightarrow \infty} C(w|w^n) = 1;$$

C is **inductive** if it is inductive about every w . Note that, while Weak Inductivism is compatible with Skepticism, Inductivism does entail Non-Skepticism (assuming countable additivity).

2. *Strong Dogmatism*. Say C is **strongly dogmatic about** w if $C(\{w\}) > 0$. Note that, since Ω is uncountable, C must assign zero credence to some world in Ω , on pain of violating finite additivity. So C is strongly dogmatic about w only if it is not weakly regular. In particular, $C(w|w \vee v) = 1$ for any v with $C(\{v\}) = 0$.

It is not hard to see that, if C is strongly dogmatic about w , then it is non-skeptical about w . Pomatto and Sandroni (2018, Proposition 1) prove a stronger statement: if C is countably additive (as an unconditional credence function), then strong dogmatism about w implies inductivism about w . Moreover, the assumption of countable additivity is in some sense necessary: there is a strictly finitely additive C that is strongly dogmatic and non-skeptical about w , but not inductive about w .¹ Below is a diagram that summarizes the logical relationship amongst various notions (dashed line indicates logical compatibility):

¹In particular, for every n and $w \in A$, $0 < C(w|w^n) < \epsilon$ for some fixed $\epsilon > 0$.



Appendix C

Chapter 4

Fact 1. *If f is E -available and E -evidentialist, then f is self-stable if and only if E is luminous.*

Proof. Suppose f is E -available and E -evidentialist. Then for all w , $[E = E(w)] \subseteq [f = f(w)]$ and $f(w)(E(w)) = 1$. The “only if”-direction is then straightforward: If E is luminous, then $E(w) \subseteq [E = E(w)]$, which means that $E(w) \subseteq [f = f(w)]$. So $f(w)(f = f(w)) = 1$, i.e. f is self-luminous, and therefore self-stable.

For the “if”-direction: Suppose f is self-stable. Since $E(w)$ is the strongest proposition to which $f(w)$ assigns credence 1, if $E(v) \neq E(w)$, then $f(v) \neq f(w)$ (for otherwise $f(v)(E(v) \& E(w))$ would equal 1 despite $E(v) \& E(w)$ being strictly stronger than $E(v)$). Hence $[E = E(w)] = [f = f(w)]$. On the other hand, since f is self-stable, $f(w)(f = f(w)) = 1$ for all w . So $f(w)(E(w) \& [f = f(w)]) = 1$, which means that $E(w) \subseteq [f = f(w)] = [E = E(w)]$. so E is luminous. $\square$

Fact 2. *An updating plan f is self-stable if and only if it is immodest.*

Proof. ($\Rightarrow$) Suppose f is self-stable. Let f' be a plan that is at least as available as f . Let $\mathcal{E}_f = \{E_1, \dots, E_n\}$ (recall that one of the assumptions we made is that W is finite). For each $i \in \{1, \dots, n\}$ and $w \in E_i$, let

$f(w) = P_i$ and $f'(w) = Q_i$ (since f' is at least as available, some of the Q_i 's may be identical). Then

$$\begin{aligned}
\mathbb{E}_f[\mathcal{A}(f')] &= \sum_w f(w)(w) \mathcal{A}(f'(w), w) \\
&= \sum_i P_i(E_i) \sum_w P_i(w|E_i) \mathcal{A}(f'(w), w) \\
&= \sum_i \sum_w P_i(w) \mathcal{A}(Q_i, w) \\
&\leq \sum_i \sum_w P_i(w) \mathcal{A}(P_i, w) \\
&= \mathbb{E}_f[\mathcal{A}(f)]
\end{aligned}$$

with equality if and only if $P_i = Q_i$ for all i .

($\Leftarrow$) Suppose f is not self-stable, i.e. there is w such that $f(w) \neq f(w)(\cdot|f = f(w))$. Let $f'(w) = f(w)(\cdot|f = f(w))$. Note that f' is at least as available as f , since $f(w) = f(v) = P$ implies that

$$f'(w) = f(w)(\cdot|f = f(w)) = P(\cdot|f = P) = f(v)(\cdot|f = f(v)) = f'(v).$$

As before, let $\mathcal{E}_f = \{E_1, \dots, E_n\}$ and P_i, Q_i be the value of f and f' at states in E_i respectively. Note that $P_i(\cdot|E_i) = Q_i$. Then

$$\begin{aligned}
\mathbb{E}_f[\mathcal{A}(f')] &= \sum_w f(w)(w) \mathcal{A}(f'(w), w) \\
&= \sum_i P_i(E_i) \sum_{w \in E_i} P_i(w|E_i) \mathcal{A}(f'(w), w) \\
&= \sum_i P_i(E_i) \sum_w Q_i(w) \mathcal{A}(Q_i, w) \\
&\geq \sum_i P_i(E_i) \sum_w Q_i(w) \mathcal{A}(P_i, w) \\
&= \mathbb{E}_f[\mathcal{A}(f)]
\end{aligned}$$

with equality if and only if $P_i = Q_i$ for all i . □

Fact 3. *If C defers to f , then f is self-stable.*

Proof. Suppose C defers to f , i.e. for every $w \in W$, $f(w) = C(\cdot|f = f(w))$. Then in particular

$$f(w)(f = f(w)) = C(f = f(w)|f = f(w)) = 1.$$

So f is self-luminous and therefore self-stable. □

Fact 4. *If f is $\mathbf{E}$ -available, $\mathbf{E}$ -evidentialist and self-stable, then f is self-reflective only if $\mathbf{E}$ is factive. The converse is true if f is the BCOND-plan.*

Proof. Let f be an updating plan that satisfies the three conditions. Suppose f is self-reflective, i.e. $f(w) = C(.|f = f(w))$. Then by $\mathbf{E}$ -evidentialism, $f(w)(E(w)) = C(E(w)|f = f(w)) = 1$. By regularity, $[f = f(w)] \subset E(w)$ for every w . In particular, $w \in [f = f(w)] \subset E(w)$ for every w . So $\mathbf{E}$ is factive. Conversely, suppose $\mathbf{E}$ is factive and f is the BCOND-plan. Since f is $\mathbf{E}$ -available and $\mathbf{E}$ -evidentialist, by Fact [cite], $\mathbf{E}$ is luminous if f is self-stable. It follows that $E(w) = [E = E(w)]$. Clearly $[E = E(w)] \subseteq [f = f(w)]$; the other direction also holds under the assumption of regularity. So $f(w) = C(.|E(w)) = C(.|E = E(w)) = C(.|f = f(w))$. □

Appendix D

Chapter 5

Fact 1. Fix a regular prior C and a strictly proper accuracy score $\mathcal{A}$. For any updating plan f ,

$$\mathbb{E}_C[\mathcal{A}(f^*)] \geq \mathbb{E}_C[\mathcal{A}(f)].$$

with equality if and only if C defers to f (i.e. $f = f^*$).

Proof. Let $\{P_1, \dots, P_n\}$ be the range of f and $Q_i = C(\cdot | f = P_i)$. Then

$$\begin{aligned} \mathbb{E}_C[\mathcal{A}(f)] &= \sum_w C(w) \mathcal{A}(f(w), w) \\ &= \sum_i C(f = P_i) \sum_w C(w | f = P_i) \mathcal{A}(P_i, w) \\ &= \sum_i C(f = P_i) \sum_w Q_i(w) \mathcal{A}(P_i, w) \\ &\leq \sum_i C(f^* = Q_i) \sum_w Q_i(w) \mathcal{A}(Q_i, w) \\ &= \mathbb{E}_C[\mathcal{A}(f^*)] \end{aligned}$$

with equality if and only if $f = f^*$. □

Fact 2. Fix a strictly proper accuracy score $\mathcal{A}$. If C defers to f , then (C, f) is not strongly accuracy-dominated by another pair (C', f') , with f' at least as available as f . Suppose $\mathcal{A}$ is in addition extensional, separable and continuous (e.g. $\mathcal{A}$ is the Brier score), then (C, f) is strongly accuracy-dominated by another pair (C', f') such that C' defers to f' and f' is at least as available as f .

Proof. For the first claim: Suppose towards a contradiction that C defers to f but (C, f) is strongly accuracy-

dominated by another pair (C', f') , with f' at least as available as f . For simplicity, let $\mathcal{A}((C, f), w)$ denote the sum of $\mathcal{A}(C, w)$ and $\mathcal{A}(f(w), w)$. Note that

$$\mathbb{E}_C[\mathcal{A}(C, f)] = \mathbb{E}_C[\mathcal{A}(C)] + \mathbb{E}_C[\mathcal{A}(f)]$$

and so is $\mathbb{E}_C[\mathcal{A}(C', f')]$.

Since (C', f') strongly accuracy-dominates (C, f) , we have

$$\mathbb{E}_C[\mathcal{A}(C', f')] > \mathbb{E}_C[\mathcal{A}(C, f)]$$

On the other hand, by Strict Propriety and Fact 5 (see below)

$$\mathbb{E}_C[\mathcal{A}(C, f)] = \mathbb{E}_C[\mathcal{A}(C)] + \mathbb{E}_C[\mathcal{A}(f)] \geq \mathbb{E}_C[\mathcal{A}(C')] + \mathbb{E}_C[\mathcal{A}(f')].$$

A contradiction.

On the other hanve Briggs and Pettigrew (forthcoming) prove that, if C does not defer to f , or in their terminology that f is not the conditionalizing strategy for C relative to the partition $\mathcal{E}_f$, then given the assumptions about $\mathcal{A}$, there is (C', f') where f' is the conditionalizing strategy for C' with respect to $\mathcal{E}_f$ (which means that f' is at least as available as f) such that $\mathcal{A}(C', f')$ strongly dominates $\mathcal{A}(C, f)$. $\square$

Fact 3. *A pair (C, f) is Dutch-bookable if and only if C defers to f .*

Proof. $(\Leftarrow)$ Suppose $f \neq f^*$, i.e. there is $w \in W$ and $A \subseteq W$ such that $f(w) = P$ and $P(A) \neq C(A|f = P)$. Without loss of generality, suppose $C(A|f = P) = a < P(A) = b$. Since C is regular, $C(f = P) = c > 0$. Consider $(X + Y, S)$ given by

$$S_w(v) = \begin{cases} 1 - b & \text{if } f(w) = P, v \in A \\ -b & \text{if } f(w) = P, v \notin A \\ 0 & \text{otherwise} \end{cases}$$

and

$$X(w) = \begin{cases} a - 1 & \text{if } f(w) = P, v \in A \\ a & \text{if } f(w) = P, v \notin A \\ 0 & \text{otherwise,} \end{cases} \quad Y(w) = \begin{cases} (b - a)(1 - c) & \text{if } f(w) = P \\ -(b - a)c & \text{if } f(w) \neq P. \end{cases}$$

Note that $\mathbb{E}_{f(w)}[S_w] = 0$ for every w and $\mathbb{E}_C[X] = \mathbb{E}_C[Y] = 0$. So $(X + Y, S)$ is acceptable for (C, f) . On

the other hand,

$$X(w) + Y(w) + S_w(w) = c(a - b) < 0.$$

So $(X + Y, S)$ is a Dutch book against (C, f) .

($\Rightarrow$) Suppose $f = f^*$. Then in particular $\mathbb{E}_C(\cdot | f = P) = \mathbb{E}_P$. Suppose towards a contradiction that (X, S) is a Dutch book against (C, f) . Then $\mathbb{E}_C[X] \geq 0$ and $\mathbb{E}_{f(w)}[S_w] \geq 0$ for all $w \in W$. However, it follows that

$$\mathbb{E}_C[X + S] = \mathbb{E}_C[X] + \sum_i C(f = P_i) \mathbb{E}_i[S | f = P_i] = \mathbb{E}_C[X] + \sum_i C(f = P_i) \mathbb{E}_i[S_i] \geq 0$$

which contradicts the assumption that (X, S) incurs a sure loss. $\square$

Fact 4. Fix C and a set of bets $\mathcal{O}$. Then for any updating plan f ,

$$\mathbb{E}_C[S_{f^*}] \geq \mathbb{E}_C[S_f].$$

with equality if and only if C defers to f .

Proof. Let $\{P_1, \dots, P_n\}$ be the range of f and $Q_i = C(\cdot | f = P_i)$. Then

$$\begin{aligned} \mathbb{E}_C[S_f] &= \sum_w C(w) S_f(w) \\ &= \sum_i C(f = P_i) \sum_w C(w | f = P_i) O_f(w)(w) \\ &= \sum_i C(f = P_i) \sum_w Q_i(w) O_f(w)(w) \end{aligned}$$

Note that, for each i , $Q_i = f^*(v)$ for some v . Since the agent acts to maximize expected utility,

$$\mathbb{E}_{f^*(v)}[S_{f^*}] = \sum_w Q_i(w) S_{f^*}(v)(w) = \sum_w Q_i(w) S_{f^*}(w)(w) \geq \sum_w Q_i(w) S_f(w)(w) = \mathbb{E}_{f(w)}[O_f].$$

Therefore, $\mathbb{E}_C[S_f] \leq \mathbb{E}_C[S_{f^*}]$, with equality if and only if $S_f = S_{f^*}$. $\square$

Fact 5. Fix a regular prior C and a strictly proper accuracy score $\mathcal{A}$. For any two self-reflective updating plans f and f' , if $f \preceq f'$, then

$$\mathbb{E}_C[\mathcal{A}(f)] \geq \mathbb{E}_C[\mathcal{A}(f')].$$

with equality if and only if they are equally sensitive.

Proof. Suppose $f \preceq f'$, i.e. $f(w) = f(v) \Rightarrow f'(w) = f'(v)$. Let $\{P_i\}_{i \geq 1}$ be the possible ranges of f and $E_i = [f = P_i]$. Let Q_i be the value of f' at worlds in E_i (some of the Q_i 's may repeat). Then

$$\begin{aligned}
\mathbb{E}_C[\mathcal{A}(f)] &= \sum_w C(w) \mathcal{A}(f(w), w) \\
&= \sum_i C(f = P_i) \sum_w C(w|f = P_i) \mathcal{A}(P_i, w) \\
&= \sum_i C(f = P_i) \sum_w P_i(w) \mathcal{A}(P_i, w) \\
&\geq \sum_i C(f = P_i) \sum_w P_i(w) \mathcal{A}(Q_i, w) \\
&= \mathbb{E}_C[\mathcal{A}(f')].
\end{aligned}$$

□

It remains to prove Result II. Fix a probabilistic prior $C \in \Delta(W)$ and two self-reflective updating plans f and f' (relative to C). Say (C, f') is **$\mathcal{E}_f$ -accuracy dominated** if there is $C' \in \Delta(W)$ and f'' at least as available as f such that (C, f') is strongly accuracy-dominated by (C', f'') . Intuitively, (C, f') being $\mathcal{E}_f$ -accuracy-dominated means that: for an agent who can follow updating plan f (i.e. she is sensitive to the conditions under which f gives different recommendations), the combination of (C, f') is inferior to the combination of a different prior and an alternative updating plan that she can follow. Clearly, if $f \prec f'$ and both f and f' are self-reflective, then f' is not a conditionalizing strategy for C with respect to $\mathcal{E}_f$, which means it is accuracy-dominated. On the other hand, if $f' \preceq f$, then since $\preceq$ is transitive, every $\mathcal{E}_f$ -available updating plan is also $\mathcal{E}_{f'}$ -available. So f' is not $\mathcal{E}_f$ -accuracy-dominated if it is not $\mathcal{E}_{f'}$ -accuracy-dominated. But f' is the conditionalizing strategy for C with respect to $\mathcal{E}_{f'}$ (by self-reflectivity). So it is not $\mathcal{E}_{f'}$ -accuracy-dominated.

Basically the ideas behind “ $\mathcal{E}_f$ -Dutch-bookability” and “optimal $\mathcal{E}_f$ -choosing strategy” are the same, so I shall introduce the definitions but skip the details of their proofs. Again let f, f' be two updating plans. We say $(\mathcal{O}_1, \mathcal{O}_2)$ is a $\mathcal{E}_f$ -Dutch book if there is an acceptance strategy (X, S) , where $f(w) = f(v) \Rightarrow S(w) = S(v)$ (i.e. when the agent accepts the same bet is constrained by f), that is acceptable but incurs the agent a sure loss. Similarly, fix an option set $\mathcal{O}$. We say S is an $\mathcal{E}_f$ -choosing strategy if $f(w) = f(v) \Rightarrow S(w) = S(v)$. Intuitively, what the results say is that: if f and f' are both self-reflective with $f \prec f'$ and when the agent's future self takes the same bet is only constrained by f , then the agent is better off by choosing in accord with f rather than f' . In other words, there shouldn't be a “mismatch” between the sensitivity of her choosing dispositions, and the sensitivity of her updating plans.