Bayesian Learning, Meager Sets and Countably Additive Probabilities: replies to G.Belot and A.Elga

Jessi Cisewski, Joseph B. Kadane, Mark J. Schervish, Teddy Seidenfeld, and Rafael Stern

Abstract

We respond to G.Belot's (2013) criticism that Bayesian theory is *epistemologically immodest* in the way it permits/requires assigning probability 0 to topologically "typical" (i.e. comeager) sets. We establish that Belot's topological conditions are epistemologically excessive for probability models of sequences of random variables. We offer a rival analysis: Differing scientific opinions should be responsive to new facts as a way to resolve their disputes. Using a result of Blackwell and Dubins (1962), we explain how amenability to new evidence may serve as the basis for resolving conflicts and help to identify epistemologically immodest Bayesian credal states. Also we assess A. Elga's (2016) rebuttal to Belot's argument.

Key Words: Bayesian learning, Bayesian consensus, comeager set, countable additivity, finitely additive probability, foregone conclusions, merging of posterior probabilities, non-conglomerability.

Bayesian Learning, Meager Sets, and Countably Additive Probabilities:
replies to G. Belot and A. Elga.

1. Introduction. Consider this laudable cognitive goal: Given a partition of rival hypotheses and an appropriate increasing sequence of shared statistical evidence, different investigators' conditional probabilities all approach 1 for the one true hypothesis in the partition. Savage (1954, Sections 3.6 and 4.6) offers basic results from the Personalist theory of probability about how this may happen across several investigators who hold non-extreme views over a common finite partition of statistical hypotheses.[1] Savage shows that – using a (finitely additive) weak-law of large numbers – given increasing statistical evidence, different non-extreme Personalists' conditional probabilities become ever more concentrated on the same one true statistical hypothesis from among a finite partition of rival statistical hypotheses.[2,3] Thus, he addresses both issues of certainty and consensus with increasing shared evidence.

Savage (1954, p. 50) notes that this result may be extended in several ways by adapting the central limit theorem, the strong law of large numbers, and the law of the iterated logarithm to sequences of conditional probabilities generated by an increasing sequence

---

[1] Say that an investigator with degree of belief represented by a probability P(·) holds a non-extreme opinion about an event E if $0 < P(E) < 1$.

[2] Savage's (1954) axiomatic theory of preference, based on postulates P1-P7, is about an idealized Bayesian agent's *static* preference relation over pairs of acts – preferences at one time in the idealized agent's life. The theory of personal probability and conditional probability that follows from P1-P7 is about an idealized agent's epistemic state at that one time: her/his degrees of belief and conditional degrees of belief at that one time. More familiar in the Bayesian literature is a *dynamic* Bayesian rule where conditional probability models the idealized agent's changing beliefs over time, when new evidence is accepted. For details on differences between the *static* and *dynamic* use of conditional probability, see Levi (1984, §4.3).

[3] We illustrate the weak and strong laws of large numbers for independent, identically distributed Bernoulli trials. Let X be a Bernoulli variable with possible values {0, 1}, where $P(X = 1) = p$, for some $0 \leq p \leq 1$. Let X_i ($i = 1, 2, \ldots$) be a denumerable sequence of Bernoulli variables, with a common parameter $P(X_i = 1) = p$ and where trials are independent. Independence is expressed as follows. For each $n = 1, 2, \ldots$, let $S_n = \sum_{i=1}^{n} X_n$. Then $P(X_1 = x_1, \ldots, X_n = x_n) = p^{S_n} \times (1-p)^{(n-S_n)}$.

The weak-law of large numbers for *iid* Bernoulli trials asserts that for each $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|S_n/n - p| > \varepsilon) = 0.$$

The strong-law of large numbers asserts that for each $\varepsilon > 0$,

$$P(\lim_{n\to\infty} |S_n/n - p| > \varepsilon) = 0.$$

If P is countably additive, the strong-law version entails the weak-law version.

of random samples. The last two of these laws require stronger assumptions than are needed for the (finitely additive) weak-law convergence result that Savage presents. In particular, they require the assumption that (conditional) probabilities are countably additive.

Consider an uncountably infinite probability space generated by increasing finite sequences of observable random variables. Rather than requiring that different agents hold non-extreme views about all possible events, which is mathematically impossible with real-valued probabilities, instead require that they agree with each other about which events in this infinite space have probability 0. They share in a family of *mutually absolutely continuous* probability distributions. If the agents' personal probabilities over these infinite spaces also are countably additive, then strong-law convergence theorems yield strengthened results about asymptotic consensus and certainty with increasing shared evidence. We make these remarks precise and discuss several of these strengthened results in Section 4.

In a recent paper critical about the methodological significance of some of the strengthened versions of Savage's convergence result, G. Belot arrives at the following harsh conclusion:

> The truth concerning Bayesian convergence-to-the truth results is significantly worse than has been generally allowed – they constitute a real liability for Bayesianism by forbidding a reasonable epistemological modesty.
> [Belot, 2013] p. 502

We argue that this verdict is mistaken.

What makes a (coherent) Bayesian credal state *over-confident* and lacking in *epistemological modesty*? Does the Bayesian position generally forbid "a reasonable epistemological modesty," as Belot intimates? These questions are both interesting and imprecise. There is no doubting that the standard of mere Bayesian *coherence* of a credal state falls short of characterizing the set of *reasonable* credal states. To use an old and tired example, a person who thinks each morning that it is highly probable that the world ends later that afternoon does not thereby violate the technical norms of *coherence*. Basic theorems of the probability calculus, particularly those concerning the asymptotic approach to certainty, apply equally to *reasonable* and *unreasonable* coherent degrees of belief.

Belot supplements Bayesian coherence with a topological standard for what he calls a "*typical*" event, in order to help identify over-confident, epistemologically immodest credal states. When a coherent agent assigns probability 0 to a topologically large set, specifically when a probability null set is *comeager*, Belot thinks that is a warning sign of *epistemological immodesty*.[4] Such a Bayesian agent is practically certain that a topologically *typical* event does *not* occur, and then Bayesian conditioning (almost surely) preserves that certainty in the face of new evidence.

Belot (2013, Section 4) argues that the Bayesian convergence-to-the-truth results about hypotheses formulated in terms of sets of observable sequences therefore are suspicious. Those results allow the Bayesian agent to dismiss a set of sequences where the convergence-to-the-truth fails because the failure set is a "small" null set, a set of personal probability 0. Except, Belot complains, that failure set may be comeager in the topology for the observables; hence, the failure set may be "typical."

We understand this to be Belot's primary requirement.
Topological Condition #1: Do not assign probability 1 to a meager set of observables.

But also Belot argues for a more demanding standard, which we call his secondary requirement.[5]

---

[4] A topologically *meager* set is one that is a denumerable union of nowhere dense sets. A *comeager* set, or a *residual* set, is the complement of a meager set.
[5] Belot (2013, p. 488), **Remark 2**, notes that when R is meager element of a measurable space $\langle \Omega, \mathcal{B} \rangle$ then the set of probabilities that assign R probability 0 is comeager in the space of probability distributions on the same measurable space $\langle \Omega, \mathcal{B} \rangle$. This points to what we label as ***Topological Condition #2***. It is a higher-order application of Belot's idea that a topologically typical set (i.e., a comeager set) should be reflected with probability 1. For Condition #2, that reasoning is applied to typical sets of probabilities.

Topological Condition #2: If H is a (relevant) meager hypothesis in the space of sequences of observables, then H should be a null set: a set with probability 0.

In Section 2 we argue that ordinary statistical models violate Topological Condition #1 independent of concerns about Bayesian reasoning. Already, Condition #1 is inconsistent with the strong laws of large numbers, including the ergodic theorem, which are asymptotic results for unconditional probabilities. (See Oxtoby [1980, p. 85].) Also, we show that Topological Condition #2 entails radical, probabilistic *apriorism* towards observed relative frequencies that has little to do with questions about Bayesian over-confidence. In particular, this topological standard requires that with probability 1, relative frequencies for an arbitrary sequence of (logically independent) events oscillate maximally. From a Bayesian point-of-view, almost surely new evidence then leaves this extreme epistemic attitude wholly unmodified. A Bayesian agent whose credal state conforms to Condition #2 knows that she/he is practically certain never to change her/his mind that the relative frequencies for a sequence of events oscillate maximally.

Elga (2016) responds to Belot's criticism by focusing on the premise of countable additivity for probability, which is needed for the strong-law versions of Savage's convergence results. Elga's response is based on an example that purportedly shows how, using a finitely but not countably additive probability, Belot's standard for being a *humble* Bayesian credal state may be satisfied without also being burdened with the immodesty of a comeager null set. In Elga's example he argues that the associated set of data-sequences where the convergence to the truth result fails has positive probability. Elga asserts that in his example, the agent's finitely additive conditional probabilities do not (almost surely) converge to the true statistical hypothesis about limiting relative frequencies; hence, such a Bayesian agent is *humble*.

In Section 3 we dispute Elga's analysis. We argue that, contrary to Elga's assertions, the merely finitely additive probability model that he uses satisfies the almost-sure convergence-to-the-truth theorem. Nonetheless, we agree that merely finitely additive probability models open the door to cases in which the strengthened convergence result fails, as we illustrate using a variant of Elga's model. But this phenomenon of non-convergence-to-the-truth with increasing evidence, which some mere finitely additive probabilities display, does not provide a rebuttal to Belot's analysis. It does not show that such merely finitely additive probabilities are reasonable. And Elga's example does not answer for those *immodest* countably additive probabilities that do satisfy the strengthened convergence results.

In Section 4 we explain how considerations relating to both asymptotic certainty and asymptotic consensus may serve to help identify *epistemologically modest* coherent states, distinguishing them from some that rightly deserve an approbation of Bayesian hubris.

2. Meager sets versus null sets.

The particular Bayesian convergence-to-the-truth results that are the subject of Belot's complaints are formulated as probability strong-laws that hold *almost surely* or *almost everywhere*. As an instance of the probability 1 qualifier, revisit the familiar instance of the strong law of large numbers, as reported in fn. 3.

Let $<\Omega, \mathcal{B}, P>$ be the countably additive measure space generated by all finite sequences of repeated, probabilistically independent [*iid*] flips of a "fair" coin. Let 1 denote a "Heads" outcome and 0 a "Tails" outcome for each flip. Then a point $\mathbf{x}$ of Ω is a denumerable sequence of 0s and 1s, $\boldsymbol{x} = <\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots>$, with each $\boldsymbol{x}_n \in \{0, 1\}$ for $n = 1, 2, \ldots$. Let $X_n(\boldsymbol{x}) = \boldsymbol{x}_n$ designate the random variable corresponding to the outcome of the n^{th} flip of the fair coin. $\mathcal{B}$ is the Borel σ-algebra generated by "rectangular" events, those determined by specifying values for finitely many coordinates in Ω. P is the countably additive *iid* product fair-coin probability that is determined by

$$P(X_n = 1) = 1/2 \quad (n = 1, 2, \ldots)$$

and where each finite sequence of length n is equally probable,

$$P(X_1 = x_1, \ldots, X_n = x_n) = 2^{-n}.$$

Let $L^{½}$ be the set of infinite sequences of 0*s* and 1*s* with limiting frequency ½ for each of the two digits: a set belonging to $\mathcal{B}$. Specifically, let $S_n = \sum_{i=1}^{n} X_n$. Then $L^{½} = \{\omega: \lim_{n\to\infty} S_n/n = 1/2\}$. The strong-law of large numbers asserts that $P(L^{½}) = 1$. What is excused with the strong law, what is assigned probability 0, is the null set N (= $[L^{½}]^{\mathbf{c}}$) consisting of the complement to $L^{½}$ among all denumerable sequences of 0*s* and 1*s*.

The null set N is large, both in cardinality, and in category under the product topology for 2^{ω}. It is a set with cardinality equal to the cardinality of its complement, the cardinality of the continuum.[6] More to the point, in reply to Belot's analysis, when 2^{ω} is equipped with the infinite product of the discrete topology on {0, 1}, then the null set N is topologically large. N is a comeager set (Oxtoby, 1980, p. 85).[7] That is, the set $L^{½}$ is meager and so is judged "small," or not-typical. Evidently, by Condition #1, the *fair coin* model is "epistemologically immodest" with respect to the Cantor Space of denumerable sequences of possible coin flips: the space of observations.

This strong-law example should come as no surprise in the light of the following result, reported by Oxtoby (1980, T.1.6):

(*) Each non-empty interval on the real line may be partitioned into two sets, $\{N, M\}$ where N is a Lebesgue measure null set and its complement $M = N^{\mathbf{c}}$ is a meager set.

---

[6] For each $0 \leq y \leq 1$, with $y \neq ½$, N contains at least one sequence with limiting relative frequency y, and these are pairwise different sequences for different values of y.

[7] Oxtoby sketches the proof of this claim in the *Supplementary Notes* [1980, p. 99]. The claim follows from an elegant application of the Banach-Mazur Game. Belot's (2013, p. 498) Remark 5, fn. 41, adapts Oxtoby's argument to show the following

Consider a point $\boldsymbol{x}$ in Cantor Space. A prior P is "open minded" with respect to the hypothesis $\boldsymbol{x}$ provided that, given any finite initial segment of $\boldsymbol{x}$, $(x_1, \ldots, x_m)$, there is a finite continuation $(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)$ where $P(\{x\} \mid (x_1, \ldots, x_n)) > .5$, and there exists some other finite continuation of $(x_1, \ldots, x_m)$, $(x_1, \ldots, x_{n'})$ where $P(\{x\} \mid (x_1, \ldots, x_{n'})) < .5$. Say that hypothesis x *flummoxes* prior P provided that, for infinitely many values of n, $P(\{\boldsymbol{x}\} \mid (x_1, \ldots, x_n)) > .50$ and for infinitely many values of n, $P(\{\boldsymbol{x}\} \mid (x_1, \ldots, x_n)) <. 50$. Then, the set of sequences in Cantor Space (i.e., the set of hypotheses) that *flummoxes* an open minded P is comeager in the infinite product topology on Cantor Space.

What Belot observes is a special case of Proposition (**), which we introduce and discuss on the next page.

Oxtoby generalizes (*) with his Theorem T.16.5.[8] In his illustration of T.16.5 using the strong-law of large numbers, the binary partition $\{N, L^{½}\}$ of the Cantor Space displays the direct conflict between the measure theoretic and topological senses of "small." N has probability 0 under the fair-coin model, and $L^{½}$ is a meager set in the product topology of the discrete topology on {0,1}. The tension between the two senses of "small" is not over some esoteric binary partition of the space of binary sequences, but applies to the event that the sequence has a limiting frequency 1/2.

We exemplify the general conflict encapsulated in Oxtoby's T. 16.5 with the following claim, which we use to criticize Condition #2.

Consider the space 2^{ω}, with points $\boldsymbol{x} = \langle x_1, x_2, \ldots \rangle$ of denumerable sequences of 0s and 1s, equipped with infinite product of the discrete topology on {0,1}.

Define the set of sequences $L^{<0,1>}$ consisting of those points x where,

$$lim.inf.\ \sum_{j=1}^{n} x_j/n > 0 \text{ or } lim\ sup.\ \sum_{j=1}^{n} x_j/n < 1.$$

The complement to $L^{<0,1>}$, **OM** = $[L^{<0,1>}]^{\mathbf{c}}$, is the set of binary sequences whose observed relative frequencies *Oscillate Maximally*.

(**) ***Proposition***: $L^{<0,1>}$ is a meager set, i.e., **OM** is a comeager set.

Theorem A1 of *Appendix A* establishes that sequences of (logically independent) random variables that *oscillate maximally* are comeager with respect to infinite product topologies on the sequence of random variables. The *Proposition* (**) is an instance of *Theorem A1* for binary sequences , i.e. where there are only two categories.

---

[8] Oxtoby's [1980, p. 64] Theorem 16.5 establishes that if the measure space $<\boldsymbol{X}, \mathcal{B}, P>$, satisfies

- P is nonatomic,
- $\boldsymbol{X}$ has a metrizable topology $\mathcal{T}$ with a base whose cardinality is less than the first weakly inaccessible,
- and, the σ-field $\mathcal{B}$ includes the Borel sets of $\mathcal{T}$,

then $\boldsymbol{X}$ can be partitioned into a set of P-measure 0 and a meager set.

What the *Proposition* establishes is that only extreme probability models of relative frequencies pass Topological Condition #2. That is, consider a measure space $<2^{\omega}, \mathcal{B}, P>$ where $\mathcal{B}$ includes the Borel sets from 2^{ω}, where 2^{ω} is equipped with the infinite product of the discrete topology as above. Each probability with $P(L^{<0,1>}) > 0$ produces a non-null set that is meager.

Unless a probability model P for a sequence of relative frequencies is *extreme* and assigns probability 1 to the set **OM** of the sequence of observed frequencies that oscillate maximally, then P assigns positive probability to a meager set of sequences, in violation of Condition #2. Evidently, the standard for *epistemological modesty* formalized in Topological Condition #2, which requires meager sets of relevant events be assigned probability 0, itself is *immodest* because it requires an extreme *apriori* opinion about how observed relative frequencies behave. It requires, for example, that P(**OM**) = 1. Of course, given evidence of a (non-null) observation ***o*** of observed relative frequencies, the resulting conditional probability leaves that extreme *apriori* opinion unchanged: P(**OM** | observation ***o***) = 1.

Familiar Bayesian models also violate the weaker Topological Condition #1. Consider an exchangeable probability model over 2^{ω}. Then, by de Finetti's Theorem [1937] each exchangeable probability assigns probability 1 to the set *L* of sequences with well defined limiting frequencies for 0*s* and 1*s*. That is, then $P\{\boldsymbol{x}: \textit{lim.inf.} \sum_{j=1}^{n} x_j/n = \textit{lim. sup.} \sum_{j=1}^{n} x_j/n\} = 1$. But L is a subset of $L^{<0,1>}$; hence, L is a meager set.

Belot introduces the two Topological Conditions #1 and #2 in order to make precise the sense of *epistemological immodesty* that he detects is pervasive in Bayesian theory. However, we find that each of these two conditions is excessively restrictive, not just of Bayesian theory, but of a large class of ordinary statistical models.

3. But what if probability is merely finitely additive?

Elga [2016] responds to Belot's challenge by reminding us that the strengthened convergence results for conditional probabilities use the assumption that probability is countably additive. (Again, Savage's Theorem [1954, p.] weak-law convergence result does not require more than a finitely additive probability.) When probability is defined for a measurable space, the principle of countable additivity has an equivalent form as a principle of *Continuity*.

Let A_i ($i = 1, \dots$) be a monotone sequence of (measurable) events, where $\lim_i A_i = A$, also is a (measurable) event.

Continuity $\qquad P(A) = \lim_i P(A_i)$.

When probabilities satisfy *Continuity*, the probabilities for a class $\mathcal{C}$ of events that form a field also determine uniquely the probabilities for the smallest σ-field generated by $\mathcal{C}$. (See Halmos [1950], Theorem 13A.) And if an event H belongs to that σ-field, then H can be approximated in probability by events from the field $\mathcal{C}$. Specifically, for each $\varepsilon > 0$ there exists a $C_\varepsilon \in \mathcal{C}$ such that $P([H-C_\varepsilon] \cup [C_\varepsilon-H]) < \varepsilon$. (See Halmos [1950], Theorem 13D.) This result has important consequences when H is a *tail-field* event in 2^ω.[9]

Consider, for example, the *tail-field* event $L^{½}$ in 2^ω. Then, $L^{½}$ can be approximated ever more precisely in probability by a sequence of finite-dimensional events $\{E_n: n = 1, \dots \}$, each of which is determined by a finite number of coordinates from 2^ω. Let $\{\varepsilon_n > 0: n =1, \dots \}$ with $lim_n\, \varepsilon_n = 0$. That is, we have that for each $n = 1, \dots$, $P([L^{½}- E_n] \cup [E_n - L^{½}]) < \varepsilon_n$ and each E_n depends upon only finitely many coordinates from 2^ω. When P is the product measure for *iid* fair-coin flips and $L^{½}$ is the *tail-field* event that is to be

[9] An event T belongs to the *tail-field* of 2^ω provided that, when a point $\boldsymbol{x}$ belongs to T, so too does each point $\boldsymbol{x}'$ that differs in only finitely many coordinates from $\boldsymbol{x}$. It is straightforward to verify that the set of *tail-field* events of 2^ω form a field.

approximated, then the finite dimensional events E_n may be chosen as the set of sequences with relative frequency of 1's sufficiently close to ½ through the first n trials. However, when Continuity fails, and P is merely finitely additive, then the probabilities over $\mathcal{C}$ may fail to define the probabilities over the smallest σ-field generated by $\mathcal{C}$.

For example, pick two values $0 \leq p \neq q \leq 1$. A coherent, merely finitely additive probability $P^{p,q}$ on 2^{ω} may assign values to each finite-dimensional event according to *iid* trials with constant Bernoulli probability p, but assign probabilities to the *tail-field* events according to *iid* trials with constant Bernoulli probability q. Then, the strong law of large numbers does not entail the weak-law of large numbers with the same values. While *finite* sequences of 0s and 1s follow an *iid* Bernoulli-p product law, with $P^{p,q}$ probability 1, the tail-event of the limiting relative frequency for 1s is q.

Let P be a merely finitely additive probability on the Borel σ-algebra of 2^{ω} where $P(\cdot) = [P^{p,q}(\cdot) + P^{q,p}(\cdot)] / 2$. Elga considers the case with $p = .1$ and $q = .9$. This finitely additive probability assigns probability ½ to the *tail-field* event $L^{.1}$ (the set of sequence with limiting frequency .1) and probability ½ to the *tail-field* event $L^{.9}$ (the set of sequences with limiting frequency .9). For $\boldsymbol{x} \in 2^{\omega}$, let $\mathbf{I}_{L^{.1}}(\boldsymbol{x})$ be the indicator function for the event $L^{.1}$ and $\mathbf{I}_{L^{.9}}(\boldsymbol{x})$ the indicator function for the event $L^{.9}$.
So, $P\{\boldsymbol{x}: I_{L^{.1}}(\boldsymbol{x}) + I_{L^{.9}}(\boldsymbol{x}) = 1\} = 1$. Thus, we see from the Proposition of the previous section, Elga's example stands in violation of *Condition #1*, since with P-probability 1 the sequence of coin flips has a convergent limiting relative frequency, which is a meager set. But apart from this issue, also Elga asserts that the conditional probabilities associated with the (merely) finitely additive P-distribution fail the almost-sure strong-law convergence result.

Specifically, Elga argues that $P\{\boldsymbol{x}: I_{L^{.1}}(\boldsymbol{x}) = lim_{n\to\infty} P(L^{.1} \mid X_1, \ldots, X_n)\} = 0$, which constitutes a maximal violation of the convergence-to-the-truth result. That is, he argues that P-almost surely, in the limit (as $n \to \infty$) the conditional probability $P(\cdot \mid X_1,$

…, X_n) converges to the indicator for the wrong limiting relative frequency. P-almost surely the conditional probability for the tail-event of the limiting relative frequency converges to the indicator for a false event. He argues for this conclusion as follows.

Let $\boldsymbol{x}$ be an element of the set $L^{.1}$, a sequence with limiting relative frequency .1, which is practically certain to occur *if and only if* the $P^{.9,.1}$ coin is flipped. Otherwise, with P-probability 1, a sequence $\boldsymbol{x}$ almost surely has a limiting relative frequency .9. Then for each $\varepsilon > 0$, for each $n > n_\varepsilon$, the observed sequence $\{X_1, \ldots, X_n\}$ has a relative frequency of 1s close enough to .1 so that the posterior probability satisfies, $P(L^{.9} \mid X_1, \ldots, X_n) > 1-\varepsilon$. This conditional probability assigns high value the limiting frequency .9 though the sequence that generates the observations has limiting frequency .1. Similarly, the conditional probabilities converge to the wrong tail-field event when the sequence that generates the observations has limiting frequency .9. Elga concludes that conditional probabilities from the merely finitely additive P-model do not satisfy the (almost-sure) strong-law convergence-to-the-truth results. But this analysis is misleading because it conditions on P-null events, as we now explain.

Here is how to understand Elga's merely finitely additive P-*model* in a way that satisfies the almost-sure convergence-to-the-truth result. Define the denumerable set of countably additive probabilities $\{P_n\}$ on 2^ω so that P_n is the *iid* product of a Bernouilli-p probability for the first n-coordinates and is the *iid* product of a Bernouilli-q probability for all coordinates beginning with the n+1^st^ position. Each P_n is a countably additive probability on the measurable space $\langle 2^\omega, \mathcal{B}\rangle$. Distribution P_n has a ***change point*** after the n^{th} trial. Let the change point, $N = n$, be chosen according to a purely finitely additive probability, with $P(N = n) = 0$, $n = 1, 2, \ldots$. Finally, let P be the induced (marginal) unconditional probability on the Borel σ-algebra of sequences of coin-flips, $\langle 2^\omega, \mathcal{B}\rangle$.

As required in Elga's construction, P behaves as the *iid* product of a Bernoulli-p distribution on finite dimensional sets, and as the *iid* product of a Bernoulli-q distribution on the *tail-field*.[10] P satisfies the weak-law of large numbers over finite sequences with Bernoulli parameter p and satisfies the strong-law of large numbers on the tail-field with Bernoulli parameter q. Hence, the strong-law does not entail the weak-law with the same parameter value.

Given an observed history, $h_j = \{X_1 = x_1, X_2 = x_2, \ldots, X_j = x_j\}$, the Bayesian agent in Elga's example assigns a purely finitely additive conditional probability to the distribution of the change point (N) so that with conditional probability 1, the change point is arbitrarily far off in the future. For each finite history h_j and for each $k = 1, 2, \ldots$, $P(N > k \mid h_j) = 1$. Elga's finitely additive P-model precludes learning about the change point variable, N.

So, whenever the agent observes a finite history of coin flips with observed relative frequency of heads near to .9, she/he has high posterior probability for the *tail-event* $L^{.1}$. Likewise, whenever the agent sees a finite history of coin flips with observed relative frequency of heads near to .1, she/he has high posterior probability for the *tail-event* $L^{.9}$. And since this agent is always sure, given each finite history h_j, that the change point (N) is in the distant future of the sequence of coin flips, she/he always assigns arbitrarily high posterior probability to **correctly** identifying the tail-event between $L^{.1}$ and $L^{.9}$.

For example, the agent assigns probability near 1 to observing indefinitely long finite histories that have observed relative frequencies that linger near .9 exactly when the sequence $\boldsymbol{x}$ has a limiting relative frequency of .1. This finitely additive credal state satisfies the conclusion of the almost-sure convergence-to-the-truth result: Almost surely,

[10] Elga follows Rao and Rao (1983, pp. 39-40) using the technique of Banach limits to establish the existence of a finitely additive probability corresponding to the $P^{p,q}$ distribution on repeated flips of the coin, based on the set of countably additive probabilities $\{P_n\}$. The method we use here, where the change point N is incorporated explicitly as a random variable in the finitely additive joint probability model, generates all the $P^{p,q}$ distributions over sequences of repeated flips of a coin as may be obtained with Banach limits. However, in addition it provides the added machinery needed to assess the agent's conditional credences $P(N \mid X_1, \ldots, X_n)$, which reflects also the agent's opinion about whether the sequence of coin flips passed the change point. Elga's reasoning ignores the fact that, for each $n = 1, \ldots, P(N > n \mid X_1, \ldots, X_n) = 1$, which is what the alternative analysis makes salient.

given the ***observed*** histories from a sequence $\boldsymbol{x}$, the conditional probabilities converge to the correct indicator for the tail-behavior of the relative frequencies in $\boldsymbol{x}$.

Elga's analysis to the contrary is based on having the agent consider conditional probabilities, $P(L^{.1} \mid h_n)$ at histories h_n that run beyond the change point. But with Elga's finitely additive probability P-model, the agent's credence is 0 of ever witnessing such a history. That is, Elga's argument, whose conclusion is that conditional probabilities converge to the wrong indicator function, requires the agent to condition on an event of P-probability 0. But this event is part of the null-event where a failure of the *almost sure* convergence to the truth is excused.

Apart from this peculiar merely finitely additive credal attitude that precludes learning about the change point N, there is something else unsettling about this Bayesian agent's finitely additive model for coin flips. Perhaps the following makes clearer what that problem is. Modify Elga's model so that

$$P'(\cdot) = [P^{.5,.1}(\cdot) + P^{.5,.9}(\cdot)] / 2,$$

with the change point N chosen, just as before, by a purely finitely additive probability, $P(N = n) = 0$ for $n = 1, 2, \ldots$. Then the strong-law result applies to tail-field events and, P'-almost surely, the limiting frequency for heads is either .1 or .9; also, just as in Elga's P-model. However, the two finitely additive coins, $P^{.5,.1}$ and $P^{.5,.9}$, assign the same probability to each finite history of coin flips. Letting h_n denote a specific history of length n,

$$P^{.5,.1}(h_n) = P^{.5,.9}(h_n) = 2^{-n}.$$

But then $$P'(L^{.1} \mid h_n) = P'(L^{.9} \mid h_n) = ½ = P'(L^{.1}) = P'(L^{.9}),$$

for each possible history. That is, contrary to the strengthened convergence-to-the-truth result, in this modified example the agent is certain that her/his posterior probability for either of the two *tail-event* hypotheses, $L^{.1}$ or $L^{.9}$, is stationary at the prior value .5. Under the growing finite histories from each infinite sequence of coin flips, the posterior probability moves neither towards 0 nor toward 1. Within the P'-model, there is no almost-sure strong-law convergence to the truth about these *tail-events*.

Evidently, what is unsettling about these finitely additive coin models is that the observed sequence of flips is entirely uninformative about the change point variable, N. No matter what the observed sequence, the agent's posterior distribution for N is her/his prior distribution for N, which is a purely finitely additive distribution assigning 0 probability to each possible integer value for N. It is not merely that this Bayesian agent cannot learn about the value of N from finite histories. Also, two such agents who have finitely additive coin models that disagree only on the *tail-field* parameter cannot use the shared evidence of the finite histories to induce a consensus about the *tail-field* events. (In *Appendix B* we use a merely finitely additive probability for displaying where, even in the short-run, consensus and certainty increasingly fails over all possible data-sequences.) We use these themes in Section 4 to provide a partial response to Belot's question about what distinguishes *modest* from *immodest* credal states.

4. *On asymptotic merging and consensus*.

In his [1877] paper, *The Fixation of Belief*, C.S.Peirce argues that sound methodology needs to defend a proposal for how to resolve inter-personal differences of scientific opinion. Peirce asserts that the scientific method for resolving such disputes wins over other rivals (e.g., *apriorism*, or the *method of tenacity*) by having the *Truth* (aka observable *Reality*) win out – by settling debates through an increasing sequence of observations from well designed experiments. With due irony, much of Peirce's proposal for letting *Reality* settle the intellectual dispute is embodied within personalist Bayesian methodology.[11] Here, we review some of those Bayesian resources regarding three aspects of "immodesty."

One aspect of a dogmatic (or "immodest") credal state is that it is immune to revision from the pressures of new observations. And a closely related second aspect is that two rival dogmatic positions cannot find a resolution to their epistemic conflicts through shared observations. These two suggest that a credal state can be assessed for immodesty based on:

[11] The irony is, of course, that Peirce objected to *conceptualism* (aka Personalist probabilities) because he thought that it inappropriately combined subjective and objective senses of "probability." See his [1878] *The Probability of Induction*.

(i) how large is the set of conjectures, and

(ii) how large is the community of rival opinions, and

(iii) for which sets of sequences of shared observations

does Bayesian conditionalization offer resolution to inter-personal credal conflicts by bringing the different opinions into a consensus regarding the truth.

We use as our starting point an important result due to Blackwell and Dubins (1962) about countable additive probabilities. Let $<\boldsymbol{X}, \mathcal{B}>$ be a measurable Borel product-space with the following structure. Consider a denumerable sequence of sets $\boldsymbol{X}_i$ (i = 1, …) each with an associated σ-field $\mathcal{B}_i$. Form the infinite Cartesian product $\boldsymbol{X} = \boldsymbol{X}_1 \times \ldots$ of denumerable sequences $(\boldsymbol{x}_1, \ldots) = \boldsymbol{x} \in \boldsymbol{X}$, where $\boldsymbol{x}_i \in \boldsymbol{X}_i$. That is, each $\boldsymbol{x}_i$ is an atom of its algebra $\mathcal{B}_i$. In the usual fashion, let the measurable sets in $\mathcal{B}$ be the σ-field generated by the measurable rectangles.

Definition: A *measurable rectangle* $(\boldsymbol{A}_1 \text{ x } \ldots) = \boldsymbol{A} \in \mathcal{B}$ is one where $\boldsymbol{A}_i \in \mathcal{B}_i$ and $\boldsymbol{A}_i = \boldsymbol{X}_i$ for all but finitely many i.

Blackwell and Dubins (1962) consider the idealized setting where two Bayesian agents have this same measurable space of possibilities, each with her/his own countably additive personal probability, creating the two measure spaces $<\boldsymbol{X}, \mathcal{B}, P_1>$ and $<\boldsymbol{X}, \mathcal{B}, P_2>$. Suppose that P_1 and P_2 agree on which measurable events have probability 0, and admit (countably additive) predictive distributions, $P_i(\cdot \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ (i = 1, 2), for each finite history of possible observations.[12] In order to index how much these two are in probabilistic disagreement, Blackwell and Dubins adopt a total-variation metric.

Define $\rho(\, P_1(\cdot \mid \boldsymbol{X}_1{=}\boldsymbol{x}_1, \ldots, \boldsymbol{X}_n{=}\boldsymbol{x}_n), P_2(\cdot \mid \boldsymbol{X}_1{=}\boldsymbol{x}_1, \ldots, \boldsymbol{X}_n{=}\boldsymbol{x}_n)\,) =$

$$\sup_{E \in \mathcal{B}} \left| P_1(E \mid \boldsymbol{X}_1{=}\boldsymbol{x}_1, \ldots, \boldsymbol{X}_n{=}\boldsymbol{x}_n) - P_2(E \mid \boldsymbol{X}_1{=}\boldsymbol{x}_1, \ldots, \boldsymbol{X}_n{=}\boldsymbol{x}_n) \right|$$

[12] Blackwell and Dubins use the concept of predictive distributions to mean those that admit *regular conditional distributions* with respect to the subalgebra of rectangular events. (See Breiman, 1968, p. 77 for a definition of a regular conditional distribution.) For discussion of countably additive probabilities that do not admit regular conditional distributions see Seidenfeld, Schervish, and Kadane (2001, p. 1614, Corollary 1).

The index ρ is one way to quantify the degree of consensus between the two agents who they share the same history of observations, $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. This index focuses on the greatest differences between the two agents' conditional probabilities.

Here is the related strong-law result about asymptotic consensus:

(***) [Blackwell and Dubins, 1962, Theorem 2] For $i = 1, 2$,

$$P_i\text{-almost surely, } \lim_{n\to\infty} \rho(\, P_1(\cdot \mid \boldsymbol{X}_1{=}\boldsymbol{x}_1, \ldots, \boldsymbol{X}_n{=}\boldsymbol{x}_n), P_2(\cdot \mid \boldsymbol{X}_1{=}\boldsymbol{x}_1, \ldots, \boldsymbol{X}_n{=}\boldsymbol{x}_n)\,) = 0.$$

In words, the two agents are practically certain that with increasing shared evidence their conditional probabilities will merge, in the very strong sense that the greatest differences in their conditional opinions over all measurable events in $\mathcal{B}$ will diminish to 0.

Since, for each event in the space $\mathcal{B}$, the familiar strong-law convergence-to-the truth result applies, separately, to each investigator's opinion, the added feature of merging allows a defense against the charge of individual "immodesty" by showing that two rival opinions come into agreement about the truth, almost surely, in the strong sense provided by the ρ-index. In the setting of the Blackwell-Dubins (1962) result, almost surely two such investigators agree that they can resolve all conflicts in their credal states over all elements of $\mathcal{B}$, and have their posterior probabilities almost surely concentrate on true hypotheses, by sharing increasing finite histories of observations from a sequence $\boldsymbol{x}$.

Schervish and Seidenfeld (1990, section 3) explore several variations on this theme by enlarging the set of rival credal states in order to consider larger communities than two investigators, and by relaxing the sense of merging (or consensus) that is induced by shared evidence from a common measurable space $<\boldsymbol{X}, \mathcal{B}>$. They show that, depending upon how large a set of different mutually absolutely continuous probabilities are considered, the character of the asymptotic merging varies. This is where topology plays a useful role in formalizing "immodesty."

Here, we summarize three of those results. Let $\mathcal{R}$ be the set of rival credences that conform, pairwise, to the Blackwell-Dubins conditions, above. Consider three increasing classes of such communities.

(1) If $\mathcal{R}$ is a subset of a convex set of rival credences whose extreme points are compact in the discrete topology, then all of $\mathcal{R}$ uniformly satisfies the Blackwell-Dubins merging result. That is, then merging in the sense of ρ occurs simultaneously over all of $\mathcal{R}$.

(2) If $\mathcal{R}$ is a subset of a convex set of rival credences whose extreme points are compact in the topology induced by ρ, then all that is assured is a weak-law merging. That is, if $\{P_n, Q_n\}$ is an arbitrary sequence of pairs from $\mathcal{R}$, and $R \in \mathcal{R}$ is an arbitrary credence from the set of rivals, then

$$\rho(\, P_n(\cdot|\boldsymbol{X}_1, \ \ldots, \boldsymbol{X}_n), Q_n(\cdot\ |\boldsymbol{X}_1, \ \ldots, \boldsymbol{X}_n\,) \overset{R}{\to} 0.$$

(3) And if $\mathcal{R}$ is a subset of a convex set of rival credences whose extreme points are compact in the weak-star topology induced by ρ, then not even a weak-law merging of the kind reported in (2) is assured.

Not surprising then, as the community $\mathcal{R}$ increases its membership, the kind of consensus that is assured – the version of community-wide probabilistic merging that results from shared evidence – becomes weaker. So, one way to assess the epistemological "immodesty" of a credal state formulated with respect to a measurable space $<\boldsymbol{X}, \mathcal{B}>$ is to identify the breadth of the community $\mathcal{R}$ of rival credal states that admits merging through increasing shared evidence from $\mathcal{B}$. For example, the agent who thinks each morning that it is highly probable that the world ends later that afternoon has an "immodest" attitude because there is only the isolated community of like-minded pessimists who can reconcile their views with commonplace evidence that is shared with the rest of us.

When the different opinions do not satisfy the requirement of mutual absolute continuity, the previous results do not apply directly. Instead, we modify an idea from Levi [1980,

§13.5] so that different members of a community of investigators modify their individual credences (using convex combinations of rival credal states) in order *to give other views a hearing* and, in Peircean fashion, in order to allow increasing shared evidence to resolve those differences.

Let $I = \{i_1, \ldots\}$ serve as a finite or countably infinite index set, and let $\mathcal{R} = \{P_i: i \in I\}$ represent a community of investigators, each with her/his own countably additive credence function P_i on a common measurable space $<\boldsymbol{X}, \mathcal{B}>$. It may be that, pairwise, the elements of $\mathcal{R}$ are not even mutually absolutely continuous. In order to allow new evidence to resolve differences among the investigators' credences for elements of $\mathcal{B}$ (rather than trying, e.g., to preserve common judgments of conditional credal independence between pairs of elements of $\mathcal{B}$) each member of $\mathcal{R}$ shifts to a credal state by taking a mixture of each of the investigators' credal states: a "linear pooling" of those states. Specifically, for each $i \in I$, let $\tilde{\alpha}_i = \{\alpha_{ij}: \alpha_{ij} > 0, \sum_{j=1}^{\infty} \alpha_{ij} = 1\}$ serve as a set of weights that investigator$_i$ uses to create the credal state $Q_i = \sum_{j=1}^{\infty} \alpha_{ij} P_i$ to replace P_i. It might be that for each $i \in I$, each Q_i is *self-centered* in the following sense. Let $\varepsilon > 0$. The Q_i might be *self-centered* in that $\alpha_{ii} \geq 1-\varepsilon$. Then, pairwise, the Q_i satisfy the assumptions for the Blackwell-Dubins' result (***) despite being self-centered. Depending upon the size of the community $\mathcal{R}$, using the replacement credal states $\{Q_i\}$ results (1), (2), and (3) obtain.

We conclude this discussion of probabilistic merging with a reminder that merely finitely additive probability models open the door to reasoning to a foregone conclusion [1996], in sharp contrast to the almost sure asymptotic merging and convergence-to-the-truth results associated with countably additive probability models. Key to these asymptotic results is the *Law of Iterated Expectations*.

Let X and Y be (bounded) random variables measurable with respect to a countably additive measure space $<\Omega, \mathcal{B}, P>$. With $\mathcal{E}[X]$ and $\mathcal{E}[X \mid Y = y]$ denoting, respectively, the expectation of X and the conditional expectation of X, given Y = y, then

Law of Iterated Expectations $\mathcal{E}[X] = \mathcal{E}[\ \mathcal{E}[X \mid Y]\ ]$.

As Schervish et al. established [1984], each merely finitely (and not countably) additive probability defined on a σ-field of sets fails this law even when the variable X is an indicator variable. That is, each merely finitely additive probability fails to be *conglomerable* in some denumerable partition. Specifically, with P merely finitely additive, there exists a measurable hypothesis H and denumerable partition of measurable events $\pi = \{E_i: i = 1, \ldots\}$ where

$$P(H) < infimum_{E_i \in \pi} P(H \mid E_i).$$

Then, contrary to the *Law of Iterated Expectations*, with expectations $\mathcal{E}$ over all $E_i \in \pi$,

$$P(H) < \mathcal{E}[\ P[H \mid E_i \in \pi]\ ].$$

In *Appendix B* we illustrate how two mutually absolutely continuous, merely finitely additive probabilities display *reasoning to contrary foregone conclusions* with increasing shared evidence. Because in this example the investigators' conditional probabilities for a pair of contrary hypotheses $\{H_1, H_2\}$ are non-conglomerable in the partitions of their increasing shared evidence, each investigator becomes increasingly certain of a different hypothesis as a function solely of the sample *size* of their shared evidence, regardless what those samples reveal.

The lesson of that example is this: Bayesian agents who use merely finitely additive probabilities face a trade-off between:

- added flexibility of modeling that comes with relaxing the constraint of countable additivity

versus

- added restrictions on the kinds of shared evidence necessary to achieve the desirable methodological laws about asymptotic consensus and certainty illustrated in the countably additive strong-laws.

5. Summary. Savage (1954) and Blackwell-Dubins (1962) offer important results showing that Bayesian methodology incorporates increasing shared evidence in order to temper and to resolve interpersonal disagreements about personal probabilities. We contrast interpersonal standards of asymptotic certainty and consensus with Belot's (2013) proposal to use a topological standard of "meagerness" in order to determine when a credal state is immodest.

We understand Belot's Topological Condition #1 to require that comeager sets should be assigned positive probability. Where a probability model treats a comeager set as null, that shows the model is "immodest" for dismissing a topological large set as probabilistically negligble. But, in the light of the fact that the set of sequences whose frequencies oscillate maximally is comeager, we see that all the familiar probability models violate Condition #1. We believe that, also, Belot endorses Condition #2, which requires that a *typical* set of sequences should receive a *typical* probability, i.e., a meager set should be assigned probability 0. We show that this topological standard entails extreme *apriori* credences about the behavior of observed relative frequencies. That standard mandates that, with probability 1, observed frequencies oscillate maximally in order to avoid being contained in a meager set. This standard creates its own "immodest" credal state, since (almost surely) the conditional probability from this model persists in assigning conditional probability 1 to the hypothesis that observed frequencies oscillate maximally.

We discuss Elga's (2016) reply to Belot's analysis. Elga proposes using merely finitely additive probability models in order to show that finitely additive conditional probabilities need not satisfy the asymptotic (strong-law) convergence result that Belot thinks reveals an "immodesty" in Bayesian methodology. Though we agree with Elga that the asymptotics of merely finitely additive conditional probabilities are different from those of countably additive conditional probabilities, we do not agree with Elga's analysis of his "finitely additive" P-coin model. We show that the failure of the convergence-to-the-truth result Elga describes obtains only on a set that the "finitely additive" P-coin model judges is null. Instead, we illustrate that when a merely finitely additive probability displays non-conglomerability in a partition, repeated trials from that partition will not support even the weak-law results about consensus and certainty that follow when either Savage's or Blackwell-Dubins' analysis applies. We argue that, when using the added generality afforded with merely finitely additive probabilities over countably additive probabilities, there are more restrictive conditions required in order to use the asymptotics of conditional probabilities from increasing shared evidence in order to resolve interpersonal credal disagreements.

Appendix A

In his classic discussion of measure and category, Oxtoby [1980: Theorem 1.6 (p. 4) and Theorem 16.5 (p. 64)] establishes that, quite generally, a topological space that also carries a Borel measure can be partitioned into two sets: One is a measure 0 set and the other, which is its complement, is a meager set. Here we show (*Theorem A1*) that this tension between probabilistic and topological senses of being a "small" set generalizes to sequences of random variables relative to a large class of infinite product topologies. We follow that result with a Corollary that applies this tension to Borel's Normal Number Theorem.

Let χ be a set with topology $\Im$ and Borel σ-field, $\mathcal{B}$. Let χ^∞ be the countable product set with the product topology $\Im^\infty$ and product σ-field, $\mathcal{B}^\infty$, which is also the Borel σ-field for the product topology (because it is a countable product). Let $<\mathbf{\Omega}, \mathcal{A}, \mathrm{P}>$ be a probability space, and let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random quantities such that, for each n, X_n: $\Omega \to \chi$ is $\mathcal{A}$ and $\mathcal{B}$ measurable. Define X: $\Omega \to \chi^\infty$ by $X(\omega) = <X_1(\omega), X_2(\omega), \ldots>$. Let $\mathcal{S}_X = X(\Omega)$ be the image of X, i.e., the set of sample paths of X. We denote elements of $\mathcal{S}_X$ as $y = <y_1, y_2, \ldots>$. $\mathcal{S}_X$ is a subset of χ^∞. Therefore, we endow $\mathcal{S}_X$ with the subspace topology. In the remainder of this presentation, we identify certain subsets of $\mathcal{S}_X$ as being either meager or comeager. These results depend solely on the topology for the measurable space $<\mathbf{\Omega}, \mathcal{A}>$, and not on the probability P. However, the probability P is needed in order to display the tension between the two, rival senses of being a "small" set.

In what follows we require a degree of "logical independence" between the X_n's. In particular, we need the sequence $\{X_n\}_{n=1}^{\infty}$ to be capable of moving to various places in χ^{∞} regardless of where it has been so far.

***Condition* A**: Specifically, for each j, let $B_j \in \mathcal{B}$ be a set such that B_j has nonempty interior B_j^o. Assume that for each n, for each $x = \langle x_1, \ldots, x_n \rangle \in \langle X_1, \ldots, X_n \rangle(\Omega)$, and for each j, there exists a positive integer $c(n, j, x)$ such that $\langle X_1, \ldots, X_n, X_{n+c(n,j,x)} \rangle^{-1}(\{x\} \times B_j^o) \neq \varnothing$.

Condition **A** asserts that, no matter where the sequence of random variables has been up to time n, there is a finite time, $c(n, j, x)$, after which it is possible that the sequence reaches the set B_j^o. For example, suppose that each X_n is the average of the first n in a sequence of Bernoulli random variables and that $\{\varepsilon_j\}_{j=1}^{\infty}$ is a sequence of positive real numbers whose limit is 0. If $B_j = (0, \varepsilon_j)$ for even j and $B_j = (1-\varepsilon_j, 1)$ for odd j then, independent of the particular sequence x, the longest we would have to wait to reach B_j is

$$c_{n,j} = \frac{n(1-\varepsilon_j)}{\varepsilon_j} + 1$$

in order to be sure that there is a sample path that takes us from an arbitrary initial sample path of length n to B_j by time $n+c_{n,j}$. Thus, $c_{n,j}$ is a *worst case* bound for waiting. For some $x = \langle x_1, \ldots, x_n \rangle$, the minimum $c(n, j, x)$ might be much smaller than this $c_{n,j}$. For instance, with jointly continuous random variables with strictly positive joint density in which $\langle X_1, \ldots, X_n \rangle(\Omega) = \chi^n$ for all n, then $c(n, j, x) = 1$ for all n, j, and x.

For each $y \in \mathcal{S}_X$, define $\boldsymbol{\tau}_0(y) = 0$, and for $j > 0$, define

$$\boldsymbol{\tau}_j(y) = \begin{cases} \min\{n > \tau_{j-1}(y) : y_n \in B_j\}, & \text{if the minimum is finite,} \\ \infty & \text{if not.} \end{cases}$$

Let $B = \{y \in \mathcal{S}_X : \tau_j(y) < \infty$ for all j, and

$$\text{let } A = \mathcal{S}_X \setminus B = B^{\mathbf{c}} \cap \mathcal{S}_X.$$

Note that A is the set of sample paths that visit at most finitely many of the B_j sets in the order specified. Because we do not require that the sets B_j are nested, it is possible that the sequence reaches B_k for all $k > j$ without ever reaching B_j. Or the sequence could reach B_j before reaching B_{j-1} but not after.

Theorem A1: A is a meager set.

Proof: Write $A = \cup_j C_j$, where $C_j = \{y \in \mathcal{S}_X: \tau_j(y) = \infty\}$. Then A is meager if and only if C_j is meager for every j. We prove that C_j is meager for every j by induction.

Start with $j = 1$. We have $\tau_j(y) = 1$ *if and only if* $y \in C_1 = \bigcap_{n=1}^{\infty} K_n$ where $K_n = \{y \in \mathcal{S}_X$: $y_n \in B_1^{\mathbf{c}}\}$. To see that C_1 is meager, notice that $C_1^{\mathbf{c}} = \bigcup_{n=1}^{\infty} D_n$, where

$$D_1 = \mathcal{S}_X \cap (B_1 \times \chi^{\infty}),$$

and for $n > 1$, $$D_n = \mathcal{S}_X \cap (\chi^{n\text{-}1} \times B_1 \times \chi^{\infty}).$$

Each D_n contains a nonempty sub-basic open set O_n obtained by replacing B_1 in the definition of each D_n by its interior B_1^o. So $C_1^{\mathbf{c}}$ contains the nonempty open set $O = \bigcup_{n=1}^{\infty} O_n$.

Next, we show that O is dense; hence, C_1 is meager. We verify that $O \cap E \neq \varnothing$ for every nonempty basic open set E. If E is a nonempty basic open set, then there exists an integer k and there exist nonempty open subsets $E_1, \ldots, E_k$ of χ such that

$$E = \mathcal{S}_X \cap (E_1 \times \ldots \times E_{\mathrm{k}} \times \chi^{\infty}).$$

Let $y \in E$, and let x_k be the first k coordinates of y. Then there exist points in $\mathcal{S}_X$ whose first k coordinates are x_k and whose $k + c(k, 1, x)$ coordinate lies in B_1^o. Hence,

$$O \cap E \supseteq \mathcal{S}_X \cap (E_1 \times \ldots \times E_{\mathrm{k}} \times \chi^{c(k,1,x)-1} \times B_1^o \times \chi^{\infty}) \neq \varnothing.$$

Next, for $j > 1$, assume that C_r is meager for all $r < j$. To complete the induction, we show that C_j is meager. Write

$$C_j = C_{j\text{-}1} \cup_{r=j-1}^{\infty} F_r$$

Where $\quad F_r = \{y \in \mathcal{S}_X\text{: } \boldsymbol{\tau}_{j\text{-}1}(y) = r \text{ and } y_n \in B_j^{\mathbf{c}} \text{ for all } n > r\}$.

It suffices to show that each F_r is meager.

Notice that F_r is a subset of

$$G_r = \{y \in \mathcal{S}_X\text{: } y_r \in B_j\} \cap \{y\text{: for all } n > r, y_n \in B_j^{\mathbf{c}}\}$$

It suffices to show that G_r is meager.

As in the case $j = 1$, write $\quad G_r^{\mathbf{c}} = \{y \in \mathcal{S}_X\text{: } y_r \in B_r^{\mathbf{c}}\} \cup \cup_{n=r+1}^{\infty} D_n$

where, $\quad D_n = \mathcal{S}_X \cap (\chi^{n\text{-}1} \times B_j \times \chi^{\infty})$.

Each D_n contains a nonempty sub-basic open set O_n obtained by replacing each B_j in the definition of each D_n by its interior B_j^o. So $G_r^{\mathbf{c}}$ contains a nonempty open set $O = \cup_{n=r+1}^{\infty} O_n$.

Last, we establish that O is dense; hence, G_r is meager. Reason as in the base case $j = 1$. We verify that $O \cap E \neq \emptyset$ for every nonempty basic open set E. If E is a nonempty basic open set, then there exists an integer k and there exist nonempty open subsets $E_1, \ldots, E_k$ of χ such that $E = \mathcal{S}_X \cap (E_1 \times \ldots \times E_k \times \chi^{\infty})$. Let $y \in E$, and let x_k be the first k coordinates of y. Then there exist points in $\mathcal{S}_X$ whose first k coordinates are x_k and whose $k + c(k, j, x_k)$ coordinate lies in B_j^o. Hence,

$$O \cap E \supseteq \mathcal{S}_X \cap (E_1 \times \ldots \times E_k \times \chi^{c(k,j,x_k)-1} \times B_j^o \times \chi^{\infty}) \neq \emptyset,$$

which completes the induction. $\Diamond_{T.\,A1}$

Next, return to consider the sequence $\{X_n\}_{n=1}^{\infty}$ of random variables described earlier. Suppose that each X_n is the sample average of some other sequence of random variables. That is, $X_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$, where each Y_k is finite. Assume that *Condition* **A** obtains. Namely, assume that the dependence between the Y_k is small enough so that $c(n, j, x) < \infty$, for all n, j, x. For example, assume that there exist $c < d$ with c either finite or $c = -\infty$, and either with d finite or $d = \infty$, such that for each $j > 1$ and each $y \in \langle Y_1, \ldots, Y_{j-1}\rangle(\Omega)$,

$$sup_{\omega \in A_y} Y_j(\omega) = d \text{ and } inf_{\omega \in A_y} Y_j(\omega) = c,$$

where $$A_y = \{\omega: \langle Y_1(\omega), \ldots, Y_{j-1}(\omega)\rangle = y\}.$$

Thus, *Condition* **A** obtains for a sequence of *iid* random variables. It also obtains for a sequence of random variables such that $\{Y_1, \ldots, Y_n\}$ has strictly positive joint density over $(c, d)^n$ for all n. In such a case, we could let

$$B_j = \begin{cases} [c, a] & \text{if c is finite,} \\ [-\infty, a] & \text{if } c = -\infty \end{cases}$$

where $c < a < d$. Then A contains all sample paths for which $lim\ inf_n X_n > a$ along with some sample paths for which $lim\ inf_n X_n = a$. If we repeat the construction of A for a countable collection of a_n with $a_n \downarrow c$, then the union of all of the A sets is meager. Then, the set of sample paths for which the $lim\ inf_n X_n > c$ is meager. A similar construction shows that the set of sample paths for which $lim\ sup_n X_n < d$ is meager. Hence the union of these last two sets is meager, and the sequence of sample paths along which X_n oscillates maximally is a comeager set.

Theorem A1 applies directly to the sequence $\{X_n\}_{n=1}^{\infty}$. It shows that certain sets of sample paths of this sequence are meager or comeager. If, as in the case of sample averages, each X_n is a function of $\{Y_1, \ldots, Y_n\}$, we can evaluate the category of a set of sample paths of the $\{Y_n\}_{n=1}^{\infty}$ sequence. If $\langle X_1, X_2, \ldots\rangle$ is a bicontinuous function of $\langle Y_1, Y_2, \ldots\rangle$, then the two sets of sample paths are homeomorphic. In particular, this implies that the category of a set of sample paths of one sequence will be the same as the category of the corresponding set of sample paths of the other sequence: the one is meager if and only if the other is.

In the case of sample averages, we can exhibit the bicontinuous function explicitly. To be specific, let $\chi = \Re$, and for each n, define $X_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$. Let $X = \langle X_1, X_2, \ldots\rangle$ as above, with $Y = \langle Y_1, Y_2, \ldots \rangle$. Let $\mathcal{S}_X$ and $\mathcal{S}_Y$ be the sets of sample paths of X and Y, respectively. That is, $\mathcal{S}_X = X(\Omega)$ and $\mathcal{S}_Y = Y(\Omega)$. For each $y \in \mathcal{S}_Y$, define

$$\phi(y) = (\ \ldots, \tfrac{1}{n}\textstyle\sum_{k=1}^{n} y_k, \ldots).$$

For each $x \in \mathcal{S}_X$, define

$$\varphi(x) = (x_1, 2x_2 - x_1, \ldots, nx_n - (n\text{-}1)x_{n\text{-}1}, \ldots).$$

Then, by construction, $\phi(Y) = X$ and $\varphi(X) = Y$. That is φ: $\mathcal{S}_X \rightarrow \mathcal{S}_Y$ is the inverse of ϕ: $\mathcal{S}_Y \rightarrow \mathcal{S}_X$. In order to have the category of the two sample paths to be the same, it is sufficient that both ϕ and φ are continuous. If they are continuous as functions both from and to $\Re^\infty$, then they will be continuous in their subspace topologies. It suffices to show that $\phi^{-1}(B)$ and $\varphi^{-1}(B)$ are open for each sub-basic open set B. Every sub-basic open set is of the form $B = \prod_{n=1}^{\infty} B_n$ where each $B_n = \Re$ except for one value of $n = n_0$ for which $\boldsymbol{B_{n_0}}$ is open as a subset of $\Re$. Then each of $\phi^{-1}(B)$ and $\varphi^{-1}(B)$ has the form $C \times \Re^\infty$ where C is a n_0-dimensional open subset of $\Re^{n_0}$; hence both sets are open, and we have that $\mathcal{S}_X$ is homeomorphic to $\mathcal{S}_Y$.

The *Proposition* (**) in the main text is an instance of *Theorem A1* for binary sequences.

A special case of *Theorem A1* (via *Corollary A1*, below) governs the tension between the two senses of being a "small" set in Borel's (1909) theorem about *Normal Numbers*.[13] For numbers in integer base-b, let us the write the set of its b-many digits as $\mathbf{b} = \{0, \ldots, b\text{-}1\}$. A real number is *normal* in base b if, in its decimal expansion base-b, each of the b many digits occurs with limiting frequency $1/b$.

Borel (1909) established that with respect to Lebesgue measure, almost all real numbers are normal in each of the countably many bases. As a special case of Borel's result, using binary decimal expansions of real numbers in the unit interval, [0, 1], we have an

---

[13] See chapter 1 of Billingsley (1986) for a helpful discussion of Borel's Normal Number theorem.

instance of the strong law of large numbers: Repeated flips of a "fair coin" almost surely have a limiting relative frequency of 1/2 for each of the two outcomes {0,1}.

Given a real number r written in base b, let $RF_n(r, j)$ be the relative frequency of digit j among the first n places in the base-b expansion of r.[14]

Definition: Say that a real number r has a decimal expansion in base b that *oscillates maximally* if, for each of the b-many digits $j = 0, \ldots, b\text{-}1$,

$$lim\ inf\ RF_n(r, j) = 0 \text{ and } lim\ sup.\ RF_n(r, j) = 1.$$

Here, we show that the class of real numbers that *oscillate maximally* (in each base) constitutes a comeager set.

Corollary A1: The set of real numbers that oscillate maximally in each base b is comeager in the product topology of the discrete topology on the set $\mathbf{b} = \{0, \ldots b\text{-}1\}$.

Proof: Attend to real numbers in the unit interval [0, 1] so that all significant digits occur after the decimal point. Fix a base b and consider real numbers written in their base b decimal expansion. Consider the product topology $<\boldsymbol{X}, \mathcal{B}>$, where $\boldsymbol{X} = \mathbf{b}^{\omega} = (\boldsymbol{X}_1 \times \ldots)$ with $\boldsymbol{X}_i = \mathbf{b}$ and $\mathcal{B} = (\mathcal{B}_1 \times \ldots)$ with each $\mathcal{B}_i$ equal to the $2^{\mathbf{b}}$ Boolean algebra on $\mathbf{b}$.

Let $\mathcal{G}$ be the sub-basis of rectangular sets, i.e., $(\boldsymbol{B}_1 \times \ldots \times \boldsymbol{B}_i \times \ldots) = \boldsymbol{B} \in \mathcal{B}$ where $\boldsymbol{B}_i = \mathbf{b}$ for all but finitely many values of i. *Condition* **A** obtains for relative frequencies of digits over the set of real numbers written in base-b. By the same reasoning above, where Theorem A1 is applied to the set of random variables $\{X_n\}_{n=1}^{\infty}$, also the set of numbers whose base-b decimals oscillate maximally is comeager. Last, since the countable union of meager sets is meager, the set of real numbers that oscillate maximally in each base is comeager. ◊ Corollary A1

Corollary A1 provides a generalization to Borel's *Normal Number* Theorem of Oxtoby's result about the *Strong Law of Large Numbers*.

[14] For each base b, there are countably many (rational) numbers that have dual decimal expansions. Each such representation ends with a tail of all 0s or with a tail of all the maximal digit, b-1. For the result reported here, it does not matter which decimal expansion is used.

Appendix B

Here we illustrate how non-conglomerably can lead two investigators with increasing shared evidence into reasoning to contrary foregone conclusions about a pair of different hypotheses, $\{H_1, H_2\}$. Regardless the finite sequence of shared observations, and merely as function of increasing sample size, investigator$_1$'s posterior probabilities approach 1 for hypothesis H_1 and investigator$_2$'s posterior probabilities approach 1 for the contrary hypothesis H_2.

We adapt an example of non-conglomerability due to L.Dubins [1975] to produce two mutually absolutely continuous merely finitely additive probabilities, $\{P_1, P_2\}$, where the shared evidence drives the two investigators into credal states of increased certainty about contrary hypotheses. Specifically, the example provides a binary partition $\{H_1, H_2\}$ where, for each possible finite sequence of integer-valued observations, $\boldsymbol{z_k} = \{n_1, n_2, \ldots, n_k\}$, the two sequences of conditional probabilities display increasing divergence about the pair $\{H_1, H_2\}$ as a result of reasoning to a foregone conclusion:

- $\lim_k [P_1(H_1 \mid z_k) - P_2(H_1 \mid z_k)] = 1$.
- For $i = 1, 2$, $\lim_k P_i(H_i \mid z_k) = 1$.

Define the measurable space $<X_1, \mathcal{B}_1>$ by $\boldsymbol{X}_1 = \{\{H_1, H_2\} \times \boldsymbol{N}\}$, and $\mathcal{B}_1 = 2^{\boldsymbol{X_1}}$.

Let $x \in \boldsymbol{X}_1$ with $x = \{H_i, n\}$ where: $i = 1$ or $i = 2$, and $n \in \boldsymbol{N}$. Define the random variable Z: $\boldsymbol{X}_1 \rightarrow \boldsymbol{N}$ by $Z(x) = n$.

Define two merely finitely additive probabilities P_i ($i = 1, 2$) as follows: Fix $.01 > \varepsilon > 0$.

- $P_i(H_1) = P_i(H_2) = 1/2$. So P_1 and P_2 agree on the "prior" probability of the hypotheses $\{H_1, H_2\}$.
- $P_i(\boldsymbol{Z} = n \mid H_i) = 1/2^n$. So $P_i(\cdot \mid H_i)$ is countably additive and defined for each subset $S \subseteq \boldsymbol{X}$.
- $P_i(\boldsymbol{Z} = n \mid H_{j\neq i}) = \varepsilon/2^n$. So, $P_i(\cdot \mid H_{j\neq i})$ is merely finitely additive, since

$$P_i(Z \in \boldsymbol{N} \mid H_{j\neq i}) = 1 \text{ but } \textstyle\sum_n P_i(Z = n \mid H_{j\neq i}) = \varepsilon.$$

This defines $P_i(S \mid H_{j \neq i})$ whenever $S \cap H_{j \neq i}$ is a finite/co-finite subset of $S \cap H_{j \neq i}$.

- Last, in order to define $P_i(\cdot \mid H_{j \neq i})$ for the uncountably many pairs of infinite sets $\{S, S^c\}$, where $|S \cap H_{j \neq i}| = |S^c \cap H_{j \neq i}| = \aleph_0$, choose a non-principal ultrafilter $\boldsymbol{U}$ on $\boldsymbol{N}$ and proceed as follows.

  For a set $T \subseteq \boldsymbol{N}$, let $U(T) = 1$ if $T \in \boldsymbol{U}$ and $U(T) = 0$ otherwise.

  Then, generally: $P_i(S \mid H_{j \neq i}) = \sum_{x \in S \cap H_j} P_i(\{x\} \mid H_{j \neq i}) + (1-\varepsilon)U[Z(S \cap H_j)]$.

Thus, each of the two measure spaces $<\boldsymbol{X}_1, \mathcal{B}_1, P_i>$ is *regular* (i.e., $P_i(E) = 0$ if and only if $E = \varnothing$), and each subset of $\boldsymbol{X}_1$ is P_i-measurable since $\mathcal{B}_1$ is the powerset of $\boldsymbol{X}_1$. So, trivially, the pair $\{P_1, P_2\}$ are mutually absolutely continuous with respect to the common measurable space $<\boldsymbol{X}_1, \mathcal{B}_1>$.

Each P_i fails to be conglomerable for the pair $\{H_1, H_2\}$ in the denumerably infinite partition π_Z of $\boldsymbol{X}_1$ induced by the variable Z. That is, for each possible value n of $\boldsymbol{Z}$,

$$P_i(H_i) = \tfrac{1}{2} < P_i(H_i \mid \boldsymbol{Z} = n) = 1/(1+\varepsilon)$$

And $$P_i(H_{j \neq i}) = \tfrac{1}{2} > P_i(H_{j \neq i} \mid \boldsymbol{Z} = n) = \varepsilon/(1+\varepsilon).$$

Specifically, given $Z = n$, $P_1(H_1 \mid Z = n) - P_2(H_1 \mid Z = n) = (1-\varepsilon)/(1+\varepsilon)$ whereas $P_1(H_1) - P_2(H_1) = 0$. So, the non-conglomerability of each probability P_1 and P_2 in the partition π_Z has the consequence that, given an observation $Z = n$, the two investigators' *posterior* probabilities for these two hypotheses are further from consensus than are their *priors*, which are in agreement about these two hypotheses. Moreover, their conditional probabilities given $Z = n$, move towards certainty in contrary hypotheses.[15]

[15] The ρ-discrepancy between P_1 and P_2 on the common measurable space $<\boldsymbol{X}_1, \mathcal{B}_1>$ is $(1-\varepsilon) = \rho(P_1, P_2)$. To see why, for each $\varepsilon < \delta$, define the event E_δ as follows. Choose the least integer n_δ such that $P_1[\bigcup_{n < n_\delta}(H_1 \cap n) \cup \bigcup_{n \geq n_\delta}(H_2 \cap n)] \geq 1 - \delta/2$. Then let $E_\delta = \bigcup_{n < n_\delta}(H_1 \cap n) \cup \bigcup_{n \geq n_\delta}(H_2 \cap n)$. So, $P_2(E_\delta) \leq \delta/2$. Hence, $P_1(E_\delta) - P_2(E_\delta) \geq 1-\delta$. Since P_1 and P_2 agree with each other on both marginal distributions, that is $P_1(H_1) = P_2(H_1) = \tfrac{1}{2}$, and for $T \subseteq \boldsymbol{N}$, $P_1(T) = P_2(T)$, the ρ-divergence between P_1 and P_2 is determined by the set of events $\{E_\delta\}$.

We may magnify the anomaly that new evidence decreases consensus and increases certainty in the two contrary hypotheses H_1 and H_2 by fixing the integer $\boldsymbol{k}$ and letting the shared evidence arise from $\boldsymbol{k}$-fold repeated sampling of Z – where each P_i-model treats repeated samples as conditionally identically and independently distributed given H_1 and given H_2. That is, construct the measureable space $<\mathbf{X}_k, \mathcal{B}_k>$ where $\Omega_k = \{\{H_1, H_2\} \times \boldsymbol{N}^k\}$ and $\mathcal{B}_k$ is the σ-field of the powerset of $\boldsymbol{X}_k$. Denote the corresponding two measure spaces $< \mathbf{X}_k, \mathcal{B}_k, P_i>$, with $i = 1, 2$. It is evident that $\{P_1, P_2\}$ are mutually absolutely continuous in their common measurable space $<\mathbf{X}_k, \mathcal{B}_k>$, since each is a regular measure.

Let $Z_k(x_k) = <Z(x_1), Z(x_2), \ldots, Z(x_k)>$. Suppose the shared evidence is $z_k = <n_1, n_2, \ldots, n_k>$, where for each P_i (i =1, 2), the $\boldsymbol{k}$-values of Z are conditionally *iid* given each of H_1 and H_2. Then,

$$P_i(H_i \mid z_k) = 1 / (1 + \varepsilon^k)$$

which approaches 1 as $\boldsymbol{k}$ increases.

With the sequence $z_k = <n_1, n_2, \ldots, n_k>$ of increasing shared evidence of observations, for certain these two probability models experience *decreasing consensus* about the rival hypotheses, $\{H_1, H_2\}$. The shared evidence leads each probability model into a foregone conclusion. The P_1-model becomes certain of H_1 and the P_2-model becomes certain of H_2, regardless the sequence of observations. Thus, the conditions are more restrictive with merely finitely additive probabilities than they are with countably additive probabilities even for (weak-law) consensus and/or certainty given the shared evidence from finitely many statistical observations.[16]

However, the ρ-discrepancy between the two conditional probabilities, $P_1(\cdot \mid n_1)$ and $P_2(\cdot \mid n_1)$ equals $P_1(H_1| n_1) - P_2(H_1| n_1) = (1-\varepsilon)/(1+\varepsilon) < \rho(P_1, P_2)$. So, even though there is non-conglomerability of each of P_1 and P_2 for the events $\{H_1, H_2\}$ with respect to the partition $\boldsymbol{N}$, the conditional distributions given the shared evidence $\boldsymbol{N} = n_1$ show a reduction in ρ-discrepancy compared with their unconditional distributions.

With $\boldsymbol{k}$ fixed, this same result obtains given the shared evidence $\boldsymbol{x}_k = <n_1, \ldots, n_k>$ from the measurable space $<\boldsymbol{X}_k, \mathcal{B}_k>$. That is, with respect to the two measure spaces $<\boldsymbol{X}_k, \mathcal{B}_k, P_i>$ (for i = 1, 2), as defined in the next paragraph of the main text above,

$$P_1(H_1| \boldsymbol{x}_k) - P_2(H_1| \boldsymbol{x}_k) = (1-\varepsilon^k)/(1+\varepsilon^k) < (1-\varepsilon^k) = \rho(P_1, P_2).$$

[16] This result does not contradict Savage's (1954, p. 48) equation (7) since the two investigataors have different probability models, P_1 and P_2. That is, these two do not agree on the statistical model for the observed shared data, given the finite partition, $\{H_1, H_2\}$.

References

Belot, G. (2013) *Bayesian Orgulity*. Phil. Sci. 80: 483-503.

Billingsley, P (1986) *Probability and Measure*, 2*nd* ed. New York: John Wiley.

Borel, E. (1909), Les probabilités dénombrables et leurs applications arithmétiques, Rendiconti del Circolo Matematico di Palermo **27**: 247–271.

Blackwell, D. and Dubins, L. (1962) *Merging of opinions with increasing information*. Ann. Math. Stat. **33**: 882-887.

de Finetti, B. (1937) *Foresight: Its logical laws, its subjective sources*. In H.E.Kyburg, Jr. and H.Smokler (eds.) *Studies in Subjective Probability,* 1964. New York: John Wiley, pp. 93-158. (Translated from the French version by H.E.Kyburg, Jr.)

Dubins, L.E. (1975) *Finitely additive conditional probabilities, conglomerability, and disintegrations*. Ann. Probability **3**: 89-99.

Elga, A. (2016) *Bayesian Humility*. Phil. Sci. 83: 305-323.

Kadane, J.B., Schervish, M.J., and Seidenfeld, T. (1996) *Reasoning to a Foregone Conclusion*. JASA **91**: 1228-1235.

Levi, I. (1980) *The Enterprise of Knowledge*. Cambridge: MIT Press.

Oxtoby, J.C. (1957) *The Banach-Mazur Game and Banach Category Theorem*, in Dresher, M., Tucker, A. W., & Wolfe, P. *Contributions to the Theory of Games* (Vol. 3). Princeton University Press: pp. 159-163

Oxtoby, J.C. (1980) *Measure and Category* (2nd edition). Springer-Verlag.

Peirce, C.S (1877) *The Fixation of Belief*. Popular Science Monthly.

Peirce, C.S. (1878) *The Probability of Induction*. Popular Science Monthly.

Savage, L.J. (1954) *The Foundations of Statistics*. New York: J. Wiley Pub.

Schervish, M.J., Seidenfeld, T., and Kadane, J.B. (1984) *The Extent of Non-conglomerability of finitely additive probabililties*. Z.Wahr. 66: 205-226.

Schervish, M.J. and Seidenfeld, T. (1990) *An approach to certainty and consensus with increasing evidence*. J. Statistical Planning and Inference **25**: 401-414.

Seidenfeld, T., Schervish, M.J., and Kadane, J.B. (2001) *Improper Regular Conditional Distributions*. Ann. Probability **29**: 1612-1624.