International Journal of Approximate Reasoning 151 (2022) 21–32

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Learning and total evidence with imprecise probabilities

Ruobin Gong [a,*], Joseph B. Kadane [b], Mark J. Schervish [b], Teddy Seidenfeld [b,c], Rafael B. Stern [d]

[a] Department of Statistics, Rutgers University, USA
[b] Department of Statistics, Carnegie Mellon University, USA
[c] Department of Philosophy, Carnegie Mellon University, USA
[d] Department of Statistics, University of São Paulo, Brazil

ARTICLE INFO

Article history:
Received 10 November 2021
Received in revised form 25 May 2022
Accepted 24 August 2022
Available online 8 September 2022

Keywords:
Corpus of knowledge
Forward induction
IP decision rule
Non-ignorable missing data
Sufficiency

ABSTRACT

In dynamic learning, a rational agent must revise their credence about a question of interest in accordance with the total evidence available between the earlier and later times. We discuss situations in which an observable event F that is sufficient for the total evidence can be identified, yet its probabilistic modeling cannot be performed in a precise manner. The agent may employ imprecise (IP) models of reasoning to account for the identified sufficient event, and perform change of credence or sequential decisions accordingly. Our proposal is illustrated with four case studies: the classic Monty Hall problem, statistical inference with non-ignorable missing data, frequentist hypothesis testing, and the use of forward induction in a two-person sequential game.

1. Introduction

Let $Cr_t(\cdot)$ be an unconditional probability function, called a *rational Credence function* that depicts some idealized agent's uncertainty at time t. Carnap's *Principle of Total Evidence* [2] requires that

$$Cr_t(\cdot) = Cred(\cdot \mid A_t),$$

where $Cred(\cdot \mid \cdot)$ is a conditional probability function and A_t is all the observational knowledge that the agent knows at time t. This implies the Bayesian rule of temporal updating, that if between an earlier time t_1 and a later time t_2 the agent's total knowledge changes by the observational report F, then

$$Cr_{t_2}(\cdot) = Cr_{t_1}(\cdot \mid F).$$

This Carnapian account of idealized Bayesian learning may be limiting. Sometimes, it is difficult to see how the agent's total evidence at the later time may be represented by a proposition, A_{t_2}, that reports all the *observational* knowledge accumulated prior to t_2. By observational knowledge, we mean the information is acquired by the agent, either through their own perception or the aid of measurement instruments. Observational knowledge constitutes only a portion of all knowledge available to the agent. Specifically, the agent's updated credence at the latter time t_2 should reflect not only the observational knowledge A_{t_2}, but also the epistemic fact that they have learned A_{t_2}.

* Corresponding author.
E-mail addresses: ruobin.gong@rutgers.edu (R. Gong), kadane@andrew.cmu.edu (J.B. Kadane), mark@cmu.edu (M.J. Schervish), teddy@stat.cmu.edu (T. Seidenfeld), rbstern@gmail.com (R.B. Stern).

https://doi.org/10.1016/j.ijar.2022.08.016
0888-613X/

Denote by $K_{t,M}(F)$ the event that the agent learns at time t by method M that event F obtains. The agent's *corpus of knowledge* at time t consists of the joint event

$$F \,\&\, K_{t,M}(F). \tag{1}$$

Then, the agent's credence function at t_2 should be

$$Cr_{t_2}(\cdot) = Cred\left(\cdot \mid A_{t_1} \,\&\, F \,\&\, K_{t_2,M}(F)\right),$$

which may or may not agree with the assertion

$$Cr_{t_2}(\cdot) \stackrel{?}{=} Cred\left(\cdot \mid A_{t_1} \,\&\, F\right). \tag{2}$$

Indeed, the question-marked equality in (2) will not hold, if the event F and its attainment $K_{t_2,M}(F)$ become *epistemologically entangled*. That is, the meaning of the observational report F depends on the context of its attainment, $K_{t_2,M}(F)$, in a non-trivial fashion.

As an illustration, let us recall the classic Monty Hall problem [28,27,31], which Section 2 analyzes in greater detail. The setting of the problem is as follows. A valuable prize is hidden at random behind one of three enumerated doors: A, B, or C. The other two doors hide no prize. The Contestant makes a first move by designating one of the three doors. The game's moderator Monty Hall then opens one of the other two doors to reveal an empty door. Last, the Contestant decides whether she would like to stay with the designated door as her final choice, or switching to the third and remaining closed door. She wins if her final choice door hides the prize. Without loss of generality, suppose that the Contestant designated door A as her initial door at t_1, and Monty reveals door B as empty. What is the Contestant's credence at t_2 about door A being the prize door?

The observational knowledge that the Contestant acquires between t_1 and t_2 is that door B is empty. However, one would be mistaken to think that this is the only source of evidence that influences her credence at t_2 about A being the prize door. The total evidence available to the Contestant is not only that *door* B *is empty* (the event) but also that *she learned it to be so* (the attainment), as the two are epistemologically entangled with one another.

In order to update their credence from t_1 to t_2 under epistemological entanglement, the agent will have to specify at the outset a rational credence function in relation to their corpus of knowledge (1), i.e.

$$Cred\left(F \,\&\, K_{t,M}(F)\right).$$

In the Monty Hall problem, this amounts to requiring the Contestant to model her own learning jointly with the game's outcome. In most situations, this can be a daunting requirement for two reasons. First, the rational credence function $Cred(\cdot)$ needs to be well-defined for all F, t and M. These aspects together span an enormous state space, on which probabilistic specification can be difficult, if not impossible. Second, for a general observable event F, the epistemological information of its attainment $K_{t,M}(F)$ is typically unobservable. Such is true even when granted that the agent satisfies the *KK-thesis*, i.e. whenever they know F, they know that they know it.

To circumvent the epistemological entanglement and maintain the feasibility of uncertainty reasoning using probabilities, we argue that the agent may cleverly identify an observable event that nevertheless meets the *Total Evidence* condition, i.e. some special F such that (2) holds with equality. This requires a concept of *sufficiency* of an observable with respect to a corpus of knowledge, put forward by Definition 1.

Definition 1. An event F observable between times t_1 and t_2 is said to be sufficient for $\left(F, K_{t_2}(F)\right)$ with respect to a question $\{E, E^c\}$ asked at t_2, provided that

$$Cred\left(K_{t_2}(F) \mid A_{t_1} \,\&\, F \,\&\, E\right) = Cred\left(K_{t_2}(F) \mid A_{t_1} \,\&\, F\right). \tag{3}$$

The notion of sufficiency in Definition 1 is analogous to the notion of *statistical sufficiency* in likelihood theory. If $Cred(\cdot \mid \cdot)$ describes a statistical model, F is statistically sufficient for $\left(F, K_{t_2}(F)\right)$ with respect to the question E, so long as $K_{t_2}(F)$ does not provide further statistical information about E.

The observable sufficient reduction F in Definition 1 may inhabit a richer state space than that of the underlying observational knowledge alone. By construction, F *disentangles* the observational knowledge from its attainment, and helps alleviate the modeling burden on the agent's part. Following Definition 1, Lemma 2 ensures that when F is sufficient for the total evidence gained between times t_1 and t_2, then Carnap's rule of conditionalization may be satisfied in the temporal updating of the agent's credence.

Lemma 2. *If between times t_1 and t_2 the total evidence that the agent gains is the conjunction $\left(F, K_{t_2}(F)\right)$, then*

$$Cr_{t_2}(E) = Cr_{t_1}(E \mid F)$$

if and only if F is sufficient for $\left(F, K_{t_2}(F)\right)$ with respect to the question $\{E, E^c\}$.

We discuss situations in which the agent is capable of identifying an observable event F that is sufficient for the total evidence, but cannot perform its probabilistic modeling in a precise manner. The identified event F offers more information than a mere observational report the agent can obtain between the earlier and later times. Indeed by sufficiency, F is meant to encode not only the observational report, but also the means through which the agent obtains the report. Therefore, the agent may not have a non-ambiguous probability model to account for F. We utilize imprecise probabilities to analyze an agent's change of credence as a dynamic learning process. In what follows, we illustrate our proposal using four case studies: the Monty Hall problem following the introduction (Section 2), statistical inference with non-ignorable missing data (Section 3), frequentist hypothesis testing (Section 4), and the use of forward induction in a two-person sequential game (Section 5). In each of these case studies, we demonstrate how an event and its attainment may become epistemologically entangled under a certain framing, discuss possible observable sufficient reductions to disentangle them, and showcase how IP may be employed to facilitate learning from the total evidence. Section 6 concludes with a discussion on the operational necessity of our proposal.

2. The Monty Hall problem

Continuing the Monty Hall problem described in Section 1, the Contestant knows, prior to the start of the game at t_1, that the prize was placed uniformly randomly behind one of the three doors. No further information was supplied to her at this stage of the game. Letting E denote the prize door, we have that the Contestant's credence about E at t_1 is uniform:

$$Cr_{t_1}(E) = \frac{1}{3},$$

and with her designating door $\mathtt{A}$,

$$Cr_{t_1}(E \mid \text{designate } \mathtt{A}) = \frac{1}{3},$$

for all $E \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$. Furthermore, we have that

$$Cr_{t_1}(E = \mathtt{A} \mid \text{designate } \mathtt{A}, D) = \frac{1/3}{1/3 + 1/3} = \frac{1}{2}, \tag{4}$$

for $D \in \{\mathtt{B}, \mathtt{C}\}$ denoting the door that would be revealed to the Contestant as empty. That is, the Contestant's conditional credence for door $\mathtt{A}$ to be the prize door would become $1/2$, upon knowing either door $\mathtt{B}$ or door $\mathtt{C}$ to be empty.

However, the Contestant's credence about whether $\mathtt{A}$ is the prize door at time t_2, $Cr_{t_2}(E = \mathtt{A})$, is *not* represented by the quantity in (4). As alluded to earlier, the total evidence available to the Contestant at t_2 is not just that door D is empty, but also that she learned it. As a matter of fact, Monty Hall revealed door $D \in \{\mathtt{B}, \mathtt{C}\}$ to be empty, and it is through and only through Monty's reveal that the Contestant learns door D to be empty. Therefore, the observable event F that is sufficient (in the sense of Lemma 2) for the Contestant's total evidence is

$$\text{MHReveals}(D),$$

which in the eyes of the Contestant satisfies

$$Cr_{t_1}\big(K_{t_2}(D) \text{ iff MHReveals}(D) \text{ iff } K_{t_2}(\text{MHReveals}(D))\big) = 1.$$

Having identified the observable event sufficient for her total evidence, the Contestant's credence about the prize door at time t_2 retains an element of imprecision. In the case the designated door, $\mathtt{A}$, were indeed the prized door, Monty would have the liberty to choose between either door $\mathtt{B}$ or door $\mathtt{C}$ to reveal to the Contestant, as either door would be empty. As the Contestant has no information about Monty's inclination to reveal either door when he has that choice, her conditional credence function for the joint event $(E, \text{MHReveals}(D))$ given that she designated door $\mathtt{A}$ is represented by an imprecise probability. Table 1 specifies this imprecise probability, in which E represents the true door behind which the prize stands, D the door that Monty reveals to be empty, and $x \in [0, 1]$ Monty's inclination to reveal door $\mathtt{B}$ over door $\mathtt{C}$ when he has the liberty to do both. This implies that

$$Cr_{t_1}\big(E = \mathtt{A} \mid \text{designate } \mathtt{A}, \text{MHReveals}(D = \mathtt{B})\big) = \frac{x/3}{x/3 + 1/3} \in [0, 1/2], \tag{5}$$

$$Cr_{t_1}\big(E = \mathtt{A} \mid \text{designate } \mathtt{A}, \text{MHReveals}(D = \mathtt{C})\big) = \frac{(1-x)/3}{(1-x)/3 + 1/3} \in [0, 1/2]. \tag{6}$$

Taking into account the total evidence available at t_2, the Contestant's credence $Cr_{t_2}(E = \mathtt{A})$ is equal to either (5) in case Monty revealed door $\mathtt{B}$ to her, or (6) in case Monty revealed door $\mathtt{C}$ to her. Therefore, her latter credence $Cr_{t_2}(E)$ is represented by the set of probabilities

Table 1
$Cr_{t_1}(E, \text{MHReveals}(D) \mid \text{designate A})$, where $E \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$ is the true door to the prize, and $D \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$ the door that Monty Hall Reveals to the Contestant, given that the Contestant designated door A as her initial choice.

Prize door E	Monty Hall Reveals D		
	A	B	C
A	0	$x/3$	$(1-x)/3$
B	0	0	1/3
C	0	1/3	0

$$\mathcal{P} = \{P : P(\mathtt{A}) \in [0, 1/2]\},$$

regardless of which door Monty Hall reveals to her.

It is worth noting that in this analysis, since we assume that the Contestant has no information whatsoever about Monty's inclination x, her latter credence exhibits *dilation* [26] when compared to her former credence $Cr_{t_1}(E)$. For the same E, the range of values that $Cr_{t_2}(E)$ may take strictly contains that of $Cr_{t_1}(E)$ regardless of which event in the partition of the total evidence space realizes between t_1 and t_2, that is, regardless of which (not-E) door Monty Hall reveals to the Contestant. Dilation occurs when the agent's credence function entertains the possibility of statistical independence between the focal event and the total evidence, such that whichever outcome is realized, it may either strengthen or weaken the credence about the focal event [26, Theorems 2.1–2.3]. The dilation phenomenon is further examined by [10] in the context of both the Monty Hall problem and its variant, the Three Prisoners problem [4,5].

3. Non-ignorable missing data

Suppose an experiment is designed to address questions about some feature pertaining to the N members of a population, N being potentially infinite. For each member i of the population, let X_i denote the true state of their feature. At time t_1, a simple random sample of n members of the population was surveyed. By time t_2, however, only $n_{obs} < n$ observations responded, whereas $n_{mis} = n - n_{obs}$ values are missing. Letting X_{obs} denote the collection of n_{obs} observed responses, it is widely understood that the conditional credence

$$Cr_{t_1}(\cdot \mid X_{obs})$$

is not necessarily the correct credence that the investigator should endorse at t_2. It does not take into consideration the total evidence available to the investigator, which should include the fact that a specific fraction of the sampled members did not respond.

The explicit accounting for the nonresponse requires the introduction of an additional binary observable random variable $D = (D_1, \ldots, D_n)$. If the surveyed individual i responded then $D_i = 1$, and $D_i = 0$ if they did not respond. The observed and missing observations can respectively be denoted as

$$X_{obs} = \{X_i : i = 1, \ldots, n, D_i = 1\},$$
$$X_{mis} = \{X_i : i = 1, \ldots, n, D_i = 0\},$$

and accordingly $n_{obs} = \sum_{i=1}^{n} D_i$ and $n_{mis} = \sum_{i=1}^{n}(1 - D_i)$. The investigator's total evidence at time t_2 is

$$(X_{obs}, D). \tag{7}$$

The observable event (7) is sufficient for the investigator's credence for θ if and only if

$$Cr_2(\theta) = Cr_{t_1}(\theta \mid X_{obs}, D). \tag{8}$$

The assertion (8) lies at the foundation of the missing data literature, and is key to avoiding epistemic entanglement using observable evidence. To update their credence for the scientific question of interest despite partially missing observations, the investigator must be able to supply some kind of knowledge about the nonresponse mechanism. This requirement may well be hard to satisfy. A most challenging type of nonresponse mechanism to model is the *non-ignorable* mechanism [19]. Non-ignorability refers to the case when the response probabilities depend nontrivially on the values of the missing data. By definition, then, any observed and partially missing dataset contains only limited (if any) information about the non-ignorable mechanism. This is precisely why modeling non-ignorability is difficult in practice. The investigator often must conduct post-survey coverage studies in order to gain the needed insight.

As a concrete example, suppose $X_i \in \{0, 1\}$ is a binary feature for an individual, and the investigator is interested in studying θ, the population proportion of individuals possessing the positive feature. Further suppose that the positive feature $X_i = 1$ is associated with an adverse health or social perception, e.g. that a person smokes. Therefore, an individual who possesses this positive feature is less likely to respond to the survey. (For simplicity, we do assume that if an individual

responds, then they respond truthfully.) For a real study on non-ignorability in opinion surveys concerning smoking, see [23].

The investigator posits a parametric sampling model

$$X_i \mid \theta \sim Ber(\theta),$$

and a nonresponse mechanism such that for some constant $\gamma \in [0, 1)$,

$$\begin{cases} D_i \sim Ber(\gamma) & \text{if } X_i = 1, \\ D_i = 1 & \text{if } X_i = 0. \end{cases}$$

That is, all surveyed individuals with a negative feature responded, whereas individuals with a positive feature only respond with probability $\gamma < 1$. This nonresponse mechanism is non-ignorable, because the response indicators D are dependent on the values of the missing data X_{mis}.

To proceed with the analysis, we note that the likelihood function for θ is a marginal likelihood function, integrating out the unobserved missing responses X_{mis}. Writing $s_{obs} = \sum_{i=1}^{n} X_i D_i$, the sum of observed positive responses, the likelihood function takes the form

$$\begin{aligned} P(X_{obs}, D \mid \theta) &= \sum_{X_{mis} \in \{0,1\}^{n_{mis}}} P(X_{obs}, X_{mis} \mid \theta)\, P(D \mid X_{obs}, X_{mis}) \\ &= (\theta\gamma)^{s_{obs}} (1-\theta)^{n_{obs}-s_{obs}} [\theta(1-\gamma) + (1-\theta)]^{n_{mis}}. \end{aligned}$$

For the purpose of illustration, suppose that the investigator's prior credence function $Cr_{t_1}(\theta)$ is characterized by the Beta(α, β) family of distributions, with density

$$B^{-1}(\alpha, \beta)\, \theta^{\alpha-1} (1-\theta)^{\beta-1},$$

where $B(a, b)$ is the Beta function. Writing $\alpha_{obs} = s_{obs} + \alpha$ and $\beta_{obs} = n_{obs} - s_{obs} + \beta$, by (8) we have that the investigator's posterior credence function $Cr_{t_2}(\theta)$ has density

$$f_\gamma(\theta) = c_\gamma^{-1} \theta^{\alpha_{obs}-1} (1-\theta)^{\beta_{obs}-1} [\theta(1-\gamma) + (1-\theta)]^{n_{mis}}, \tag{9}$$

where the normalizing constant

$$c_\gamma = B(\alpha_{obs}, \beta_{obs})\, \mathcal{R}((\alpha_{obs}, \beta_{obs}), (1-\gamma, 1), -n_{mis}),$$

where $\mathcal{R}(b, Z, -d)$ is Carlson's multiple hypergeometric function [6], which has been previously studied in the Bayesian modeling of censored categorical data [7,14] to represent the expectation of marginal linear combinations of Dirichlet random variables. In particular,

$$\mathcal{R}((\alpha_{obs}, \beta_{obs}), (1-\gamma, 1), -n_{mis}) = (1-\gamma)^{n_{mis}}\, {}_2F_1(-n_{mis}, \beta_{obs}; \alpha_{obs} + \beta_{obs}; \gamma/(\gamma-1)),$$

where ${}_2F_1(u_1, u_2; l_1; z)$ is the generalized hypergeometric function.

The posterior credence function $Cr_{t_2}(\theta)$ depends on the response probability γ for the positive feature. It remains for the investigator to determine what values of γ is reasonable. With γ left unspecified, the posterior credence function $Cr_{t_2}(\theta)$ in (9) induces a set of probability functions

$$\mathcal{P} = \left\{ P : P(A) = \int_A f_\gamma(\theta)\, d\theta,\ \gamma \in [0, 1) \right\}. \tag{10}$$

Fig. 1 depicts $Cr_{t_2}(\theta)$ for different values of γ, for a hypothetical sample with $(n, n_{obs}, s_{obs}) = (10, 5, 2)$. Prior credence $Cr_{t_1}(\theta)$ is uniform on $[0, 1]$, corresponding to $\alpha = \beta = 1$. Note that the triple (n, n_{obs}, s_{obs}) is a reduction of the sufficient observable event in (7), and is *minimally sufficient* for the posterior credence $Cr_{t_2}(\theta)$ in the usual sense of the phrase. As is clear from Fig. 1, the posterior credence function $Cr_{t_2}(\theta)$ exhibits large differences depending on the value of γ. If γ is small, it suggests that individuals with $X_i = 1$ are much less likely to respond. Therefore, the fact that half of the surveyed individuals did not respond should be taken as strong indication that there are more people with a positive feature $X_i = 1$ that are unobserved, and the investigator should put higher posterior credence for θ on the larger values. Whereas if γ is large, individuals with $X_i = 1$ are not much less likely to respond, and the posterior credence for θ tend towards the smaller values.

We remark that the IP treatment presented here for the case of non-ignorable missing data has close ties to the literature of partial identification in econometrics; see e.g. Chapter 1 of [22]. Indeed, the investigator's updated credence is partially identified, in the sense that the observed data (X_{obs}, D) do not provide enough discerning information to pin down $Cr_2(\theta)$ as

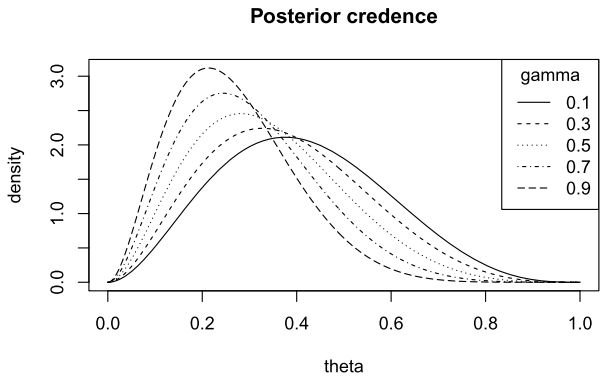

Fig. 1. Posterior credence function $Cr_{t_2}(\theta)$ in (9) for different values of γ, the individual response probability with positive feature, for a hypothetical observation with $(n, n_{obs}, s_{obs}) = (10, 5, 2)$. Prior credence $Cr_{t_1}(\theta)$ is uniform on $[0, 1]$.

a unique probability, even if a precise prior credence $Cr_1(\theta)$ is specified. The identification region of $Cr_2(\theta)$ is precisely the set of probabilities specified by (10). If the investigator would like to avoid partial identification, he or she may adopt a "full" Bayesian approach by further imposing a precise prior credence function on γ, the probability of missing the observation given a positive feature. However, since the observed data do not provide identifying information about γ, the investigator's future credences about the primary question of interest θ may be sensitive to the prior specification for γ, which calls for careful deliberation. In practice, for a chosen family of priors on γ, the investigator may employ an *empirical Bayes* approach to construct posterior estimates based on the implied marginal distribution of the data that uses, for example, plug-in estimators for θ. See [29,24] for a demonstration of small-area estimation using Bayesian hierarchical models for non-ignorable missing data, with applications to the National Crime Survey.

4. Frequentist hypothesis testing: the null hypothesis and the reference class

The previous section demonstrates how the IP construction provides a meaningful Bayesian statistical analysis in the presence of non-ignorable missing data, which are typically challenging to model precisely. In statistical inference, the virtue of IP extends beyond the context of Bayesian modeling. In this section, we discuss how one may appeal to the IP tools to identify the level of significance of an experiment in frequentist hypothesis testing, in a way that is appropriate for the total evidence available to the statistician. As the total evidence changes, many aspects of the testing procedure change as well, including the sample space, the null hypothesis, the reference class, the level of significance, as well as the final conclusion. A version of this example is discussed by Barnard [1], which we adapt here for an extended illustration.

A bag of chrysanthemum seeds are known to produce either white or purple blooms, and a statistician wishes to study the relative proportion of either color. His plan is to conduct an experiment by asking a lab scientist to sow a random selection of seeds from the bag, and he would record the colors of the flowers once they bloom. Based on the result of the experiment, the statistician would calculate the *level of significance*, defined as the maximum probability derivable under a null hypothesis H_0 of a *reference class*, a set consisting of potentially realizable experimental outcomes under the experimental design, which are considered as significant or more significant than the observed one:

$$\text{argmax}_{P \in H_0} P(R_F), \tag{11}$$

where F is the observed outcome, and the reference class R_F consists of events deemed as of matching or exceeding significance compared to F. The shorthand "$P \in H_0$" indicates that when calculating (11), every probability distribution P compatible with the null hypothesis H_0 is contemplated. The choice of P is unique when H_0 is simple, and plural when H_0 is composite. In what follows, we present three scenarios in which the total evidence available to the statistician varies in richness.

Scenario 1 Suppose a few months after the seeds are sown, the statistician observes nine blooming chrysanthemum flowers, all of which are white in color. Let n be the number of blooming chrysanthemum flowers, and r be the number (out of n) of white ones in excess over purple. The observational report can be summarized as

$$F: \quad n = 9, r = 9. \tag{12}$$

Under the null hypothesis that white and purple colors are equally likely, the statistician presumes the experiment as a Binomial trial

$$H_0: \quad \left(\frac{n+r}{2}, \frac{n-r}{2}\right) \sim Bin(n, \frac{1}{2}).$$

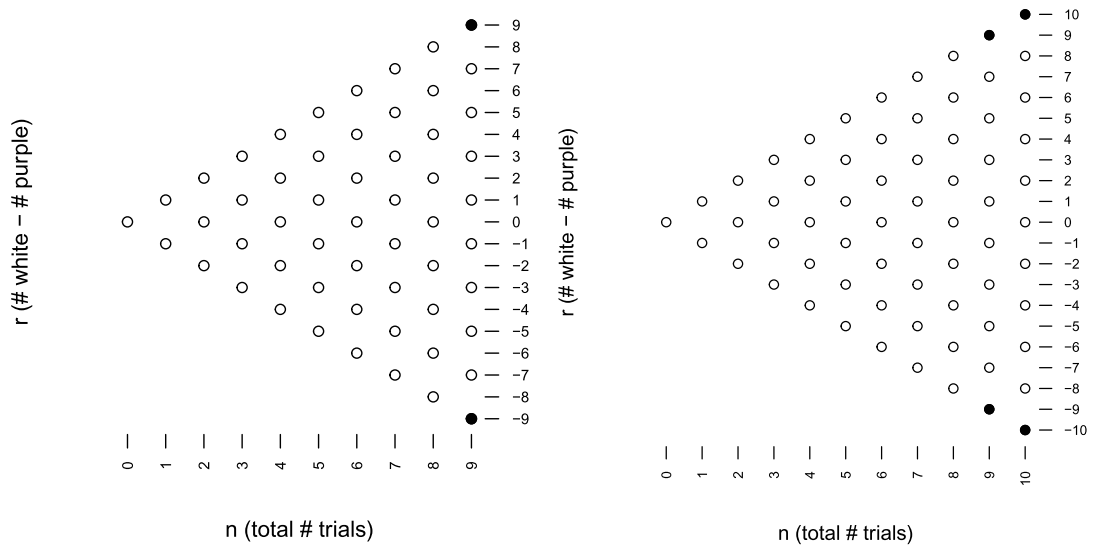

Fig. 2. Left: the sample space in Scenario 1 expressed in terms of (n, r), and the reference class R_F ((13); solid black dots); Right: the sample space in Scenarios 2 and 3, and the reference classes $R_{F'}$ and $R_{F''}$ ($R_{F'} = R_{F''}$).

Following Barnard [1], the reference class R_F of events that match or exceed the significance of the observation F in (12) is defined as

$$R_F = \left\{(\tilde{n}, \tilde{r}) : \tilde{r} \geq |r| + \min\left(0, \tilde{n} - n\right) \text{ or } \tilde{r} \leq -|r| - \max\left(0, n - \tilde{n}\right), \tilde{n} \leq n, \left|\tilde{r}\right| \leq \tilde{n}\right\} \tag{13}$$

which simplifies to $\{(9, 9), (9, -9)\}$, or obtaining all nine white flowers (as observed) or all nine purple. The left panel of Fig. 2 showcases the sample space and the reference class R_F. The reference class encodes a sense of *deviation* [see e.g. 3, Chapter 4] between the null hypothesis and the empirical law governing the observation F. In general, the choice of the reference class need not be unique, as it reflects the statistician's judgment about which outcomes are deemed more scientifically significant. Barnard [1] motivates why R_F takes the form of (13). For simplicity, we follow Barnard's choice throughout this section and omit the discussion on alternative choices of R_F, with the understanding that a different R_F may lead to different numerical conclusions from what we present here. According to (11), the corresponding level of significance of the experiment is equal to $P(R_F) = 2 \times (1/2)^9 \doteq 0.39\%$.

Scenario 2 There's more to the story than meets the eye of the statistician. What the lab scientist did not reveal to the statistician is that he sowed not nine, but in total $N = 10$ seeds, and one of them did not reach the flowering stage. The lab scientist did not fully disclose the reason why one of the seeds failed. He noted that the seed did germinate, but as a young plant was prematurely destroyed under certain unforeseen circumstances. The statistician now gathered a fuller set of evidence:

$$F': \quad n = 9, r = 9, N = 10, \text{“premature plant destruction”},$$

and must now revise both his null hypothesis and the reference class. In particular, let n_w and n_p denote the respective numbers of white and purple flowers to result from the destroyed plant. Each of n_w or n_p is unobserved, but their total $n_w + n_p = N - n$ is observed. The statistician presumes a Multinomial trial as the null hypothesis:

$$H_0': \quad \left(\frac{n+r}{2}, n_w, \frac{n-r}{2}, n_p\right) \sim \text{Multinomial}\left(N, \left(\frac{p_w}{2}, \frac{1-p_w}{2}, \frac{p_p}{2}, \frac{1-p_p}{2}\right)\right),$$

where $p_w, p_p \in [0, 1]$ are respectively the probabilities for plants bearing white and purple flowers to be destroyed. Note that since the values of p_w and p_p are under-determined, the null hypothesis H_0' is a composite one, encompassing a class of sampling distributions for the experimental outcome. Following Barnard's construction of the reference class as defined in (13), the reference class for F' becomes

$$\begin{aligned} R_{F'} = \{&(\tilde{n}, \tilde{r}, \tilde{n}_w, \tilde{n}_p) : \tilde{r} \geq |r| + \min\left(0, \tilde{n} - n\right) \text{ or } \tilde{r} \leq -|r| - \max\left(0, n - \tilde{n}\right), \\ &\tilde{n} \leq N, \left|\tilde{r}\right| \leq \tilde{n}, \tilde{n}_w + \tilde{n}_p = N - n\} \\ = \{&(9, 9, 1, 0), (9, 9, 0, 1), (9, -9, 1, 0), (9, -9, 0, 1), (10, 10, 0, 0), (10, -10, 0, 0)\}. \end{aligned}$$

Under the null hypothesis H_0', the probability of the reference class, $P(R_{F'})$, belongs to a set of probabilities

$$\left\{ \left(\frac{10p_p + 9p_w + 1}{2} \right) \left(\frac{1 - p_w}{2} \right)^9 + \left(\frac{10p_w + 9p_p + 1}{2} \right) \left(\frac{1 - p_p}{2} \right)^9 ; \ (p_w, p_p) \in [0,1]^2 \right\} \tag{14}$$

The set (14) reaches a minimum of 0 when $p_w = p_p = 1$, and a maximum of approximately 1.07% when $p_w = 0$, $p_p = 1$ or $p_p = 0$, $p_w = 1$. The latter, being the highest among all models contemplated under the null hypothesis H_0', is taken to be the level of significance of the observed experiment.

Scenario 3 Further suppose that the lab scientist finally reveals the story behind the young plant's demise. One day while running errands, he carelessly trod over it, causing its unfortunate death. This is an important piece of information. It allows the statistician to rule out the possibility of a linkage, genetic or environmental, between the destruction propensities of the plant and the color of its flower. Without knowing the precise cause of the plant's destruction, the statistician suspected that one color of the chrysanthemum is more susceptible of certain pests or diseases, hence allowing for differing destruction probabilities, p_w and p_p, in the formulation of his previous null hypothesis H_0'. Now, the statistician's total evidence can be stated as

$$F'': \quad n = 9, r = 9, N = 10, \text{“premature plant destruction due to a color-agnostic reason”}.$$

The null hypothesis is also updated by equating p_w with p_p to a common (and unknown) probability p:

$$H_0'': \quad \left(\frac{n+r}{2}, n_w, \frac{n-r}{2}, n_p \right) \sim \text{Multinomial}\left(N, \left(\frac{p}{2}, \frac{1-p}{2}, \frac{p}{2}, \frac{1-p}{2} \right) \right).$$

With the reference class remaining the same as before, $R_{F''} = R_{F'}$, its probability under the null H_0'' again belongs to a set of probabilities

$$P(R_{F''}) \in \left\{ 2 \left[10p + \frac{1-p}{2} \right] \left(\frac{1-p}{2} \right)^9 ; \ p \in [0,1] \right\} \tag{15}$$

As the common destruction probability p varies between $[0,1]$, $P(R_{F''})$ in (15) reaches a minimum of 0% (obtained when $p = 1$) and a maximum of approximately 0.24% (obtained when $p = 1/19$). Again, the latter is taken to be the level of significance of the experiment, for it is the highest derivable under the class of models implied by the null hypothesis H_0''. This last portion of the analysis agrees with Barnard's.

The three scenarios illustrated above concern the same physical experiment, but due to a difference in the total evidence, they are accompanied by three different suites of hypothesis tests. Recognizing what constitutes the total evidence impacts the construction of the test in more than one way. It may change the null hypothesis the statistician finds reasonable to entertain, as well as the realized reference class of events deemed as significant or more significant than the evidence at hand. The change in the latter is the result of a changing sample space, necessitated by the total evidence which suggests the total number of potentially observable flowers. The change in the reference class would happen, even if the observed numbers of flowers of either color remain the same, and the statistician employs the same definition of reference classes throughout the different scenarios.

Comparing across the three scenarios, we see that their respective bodies of total evidence increase in richness. F' of Scenario 2 provides information additional to F of Scenario 1 about the existence of a destroyed plant, and F'' of Scenario 3 further adds to F' by describing the condition of its destruction. The statistical models entailed by the three null hypotheses are nested, in the sense that the binomial model associated with H_0 is marginally implied by the multinomial model of H_0'', which is in turn a parameter-restricted special case to the multinomial model of H_0'. Furthermore, the analyses yielded three levels of significance, 0.39%, 1.07% and 0.24%, which are similar in value but convey distinct substantive meanings, because each corresponds to the maximum probability of a distinct reference class derivable under a distinct set of probability models under the null hypotheses. Using $\prec$ to loosely denote the respective ordering as described above, we see that none of these aspects are aligned or indicative of one other:

- Richness of total evidence: Scenario 1 $\prec$ Scenario 2 $\prec$ Scenario 3;
- Generality of null hypothesis: Scenario 1 $\prec$ Scenario 3 $\prec$ Scenario 2;
- Level of significance attained: Scenario 3 $\prec$ Scenario 1 $\prec$ Scenario 2.

Thus, richer total evidence does not entail a more general (or more specific) null hypothesis, nor does it necessarily result in a higher (or lower) assessment of level of significance.

Even though the total evidence may explicitly acknowledge the presence of unobserved data, we do not advocate for the statistical modeling of arbitrary unknown information. After all, any rational agent can only know so much, their knowledge would almost certainly be dwarfed by their ignorance. The total evidence is amenable to statistical modeling, just in case there exists an observable *sufficient* reduction in the sense of Definition 1. In the second and third scenarios examined in

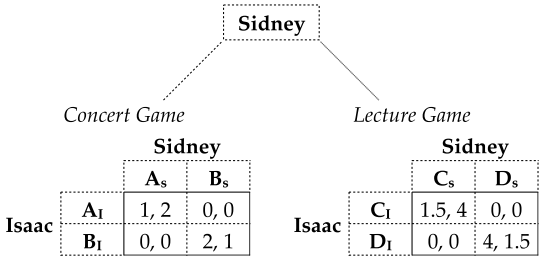

Fig. 3. A two-person, two-stage extensive form sequential game, in which one player (Sidney) chooses which subgame for both players to play at the second stage.

this section, the sufficiency of F' and F'' is exemplified by the fact that they both provide information specific enough to determine the appropriate sample space, definitive and finitely sized, as well as about the destruction mechanism to allow for a meaningful IP description. In practice, the agent may seek additional, *attainable* information which they know can aid the sufficient reduction of total evidence. As an example, to adjust for complex patterns of undercounts in the population census, the United States Census Bureau conducts *post-enumeration surveys* (PES) to measure the extent of omission using stratified random samples; see e.g. [13]. We return to this point again in Section 6.

5. Forward induction with imprecise probabilities

In this section, we consider two-person sequential games in which each player is required to specify a *grand plan* for action at each node of the game tree where the player needs to make choices. A game-theoretic criterion that dictates what kind of grand plans should be deemed as acceptable is *subgame perfection*; see e.g. [12]. Under subgame perfection, a grand plan is said to be acceptable if and only if it yields acceptable strategies within *each* sub-game of the larger game. However, since players who adhere to subgame perfection must treat each subgame as a separate game irrespective of all other aspects of the larger game, including choices that have been made by their opponents, they can violate total evidence in devising their grand plans for the game.

To take into account the total evidence, the player may endorse instead the criterion of *forward induction* [16], by recognizing and utilizing what they observe from the preceding plays that lead them into the subgame. That is, at the beginning of each subgame, the player works with the total evidence which includes not only their uncertainty model about their opponents, but also the fact that the dynamic of the sequential game has lead both of them to this particular subgame.

In addition to adopting forward induction reasoning in the sequential game, we also assume that the players employ imprecise probability models of uncertainty, represented by a set $\mathcal{P}$ of personal probability functions, rather than by a single such function, P. The literature has demonstrated the value, and indeed the necessity, of using IP models of uncertainty in game theory [e.g. 9,30,20].

We now describe the setting of the game. Two players, Sidney and Isaac, are about to play a two-stage extensive form sequential game. In the first stage of the game, Sidney chooses between their playing either the Concert Game, or playing the Lecture Game. In the Concert game, they coordinate on attending either (A) a Bruch violin concerto, played by Itzhak Perlman, or (B) a Dolly Parton concert. In the Lecture game, they coordinate on attending (C) a lecture by Chomsky, or (D) a lecture by Ellsberg. In these subgames, Sidney is Column player and Isaac is Row player. The setting of the game is illustrated in Fig. 3. The goal of both players is to arrive at a precise action plan at the end of the iterative game, while at the same time maximizing their own utility outcome to the extent possible. We pay particular attention to the interpretation of players' strategies and the utility outcomes from Isaac's point of view, whose application of forward induction highlights the impact of total evidence in guiding his subgame decision as the Row player.

In either the Concert or Lecture subgames, we allow the two players to adopt an extreme IP model that reflects maximal uncertainty about the other player's choices. That is in each subgame, each player uses the set $\mathcal{P}$ of all probabilities for what the other player might choose from among all his mixed strategies. For instance, without additional evidence, Isaac is maximally uncertain about which strategy Sidney will use in the Lecture Game. That is,

$$\mathcal{P}^{\text{Isaac}}_{\text{Lecture}}\{xA_s \oplus (1-x)\,B_s\} = \{P : 0 \le P\,(xA_s \oplus (1-x)\,B_s) \le 1\}\,,\ \forall x \in [0,1],$$

where $A \oplus B$ denotes a direct sum of the two games A and B. The two subgames each have the following three Nash equilibria pairs, two are pure and one is mixed. In the Concert Game:

- $\langle A_I, A_S\rangle$ yields utility outcome $(1,2)$;
- $\langle B_I, B_S\rangle$ yields utility outcome $(2,1)$;
- $\langle (1/3)A_I \oplus (2/3)B_I, (2/3)A_S \oplus (1/3)B_S\rangle$ yields utility outcome $(2/3, 2/3)$.

In the Lecture Game:

- $\langle C_I, C_S \rangle$ yields utility outcome $(1.5, 4)$;
- $\langle D_I, D_S \rangle$ yields utility outcome $(4, 1.5)$;
- $\langle (3/11)C_I \oplus (8/11)D_I, (8/11)C_S \oplus (8/11)D_S \rangle$ yields utility outcome $(12/11, 12/11)$.

One complication with the analysis of games represented by IP models is that there exists a variety of applicable decision rules. A different choice of rules may yield different action plans and different consequences [25,21]. In the current two-stage game, both players aim to arrive at one precise action plan, and the IP decision rule that they employ must be conducive to this goal. Thus in this example, the IP decision rule that both players employ restricts admissibility to those options that maximize minimum expectation with respect to the set $\mathcal{P}$ of probabilities, i.e. options that are Γ-maximin [8], among those options that maximize expected utility for some probability P in the set $\mathcal{P}$, i.e. options that are E-admissible [17]. This is Levi's lexicographic rule [18] that uses E-admissibility as the primary consideration, and Γ-maximin as the secondary (or *security*) consideration for admissibility. A brief discussion about the choice of IP decision rules appears at the end of this section.

Since each player has a maximally uncertain IP model for which strategy the other payer chooses in these games, each of these three Nash pairs also are pairs of E-admissible options, because each of these three strategies maximizes expected utility against the other's matching strategy. More significant, in each game, in the light of the "equalize" mixed strategy Nash equilibrium pair, *each* mixed strategy that Isaac might choose also is E-admissible against that mixed strategy for Sidney. For instance, in the Lecture Game, each mixed strategy that Isaac might play, $xC_I \oplus (1-x)D_I$, is E-admissible against Sidney's equalizer mixed strategy, $(8/11)C_S \oplus (3/11)D_S$.

Next, we turn to considerations of security maximization. In the Concert Game:

- Isaac's mixed strategy, $(2/3)A_I \oplus (1/3)B_I$ secures a minimum expectation of $2/3$, and that is the maximum security possible for Isaac among his (E-admissible) strategies.
- Likewise, by the symmetries of the game, Sidney's mixed strategy $(1/3)A_S \oplus (2/3)B_S$ secures a minimum expectation of $2/3$, and that is the maximum security possible for Sidney relative to all his (E-admissible) strategies.

In the Lecture Game, the security maximizers are

- For Isaac, $(8/11)C_I \oplus (3/11)D_I$ secures $12/11$ utility, and
- For Sidney, $(3/11)C_S \oplus (8/11)D_S$ secures $12/11$ utility.

Here is how we apply forward induction with these IP decision rules in the two-stage game between Isaac and Sidney, where Sidney plays first to choose which subgame they play. We use the following hypothetical "cheap talk" dialogue to make explicit the steps in the IP decision making.

> *Sidney*: Isaac, suppose I choose we to go to the Concert. What will you do?
> *Isaac mumbles to himself*: Well, if I saw that Sid chose the Lecture Game, that would give him an E-admissible option with a security of $12/11$. Hmm...
> *Isaac*: Then, Sid, if you choose the Concert Game (and reject the Lecture Game) you'd be signaling to me that you expect at least $12/11$ in the Concert Game. So, I'd choose to join you to hear Perlman play Bruch, and you'll get 2 units utility while I get only 1.
> *Sidney*: Very good. Let's go to the Lecture!
> *Isaac mumbles to himself*: Well, rejecting Concert means that Sid now expects at least 2 units by going to the Lecture.
> *Isaac*: Then, Sid, I see I'm stuck going to hear Chomsky with you.
> *Sidney*: Yes. But at least you'll enjoy that more than you would the Bruch!

Note, the application of forward induction illustrated in this example conforms to the conjecture that players (e.g. Isaac) can avoid the epistemic entanglement by using observable (even hypothetical) decisions from earlier in the game to fix expectations later in the game, without needing to incorporate an additional epistemic random variable for current knowledge. Indeed, in the first iteration of the game, Isaac chooses Perlman's Bruch (A_I) over his preferred Dolly Parton (B_I), because he knows that Sidney rejected altogether the Lecture subgame (and thereby a security of $12/11$), a piece of observed knowledge that precedes the current Concert subgame. Similarly, in the second iteration of the game Isaac chooses Chompsky (C_I) over his preferred Ellsberg (D_I), because he knows that Sidney rejected the Bruch concerto (and thereby a certain utility outcome of 2) in the Concert subgame, which is again observed knowledge that precedes the current Lecture subgame.

Before concluding this section, we remark on the use of Levi's lexicographic IP decision rule in this example. The primary purpose of the example is to illustrate forward induction in sequential games with ambiguity, as a means for the players to avoid the epistemic entanglement using sufficient observables. Given that both players possess vacuous knowledge about each other's strategies, this lexicographic decision rule (i.e. E-admissibility first, with Γ-maximin as secondary security) allows the players to arrive at a unique strategy. As discussed, E-admissibility alone reduces the admissible options only to the infinite number of *rationalizable* strategies. On the other hand, if both players endorse Γ-maximin without consideration

for E-admissibility, it would hinder Isaac's ability to perform forward induction and make use of Sidney's suggestion for the Concert game as evidence to guide his own choice. In Isaac's view, Sidney's strategy needs not be Nash, if he is not bound by E-admissibility.

We do not defend the lexicographic rule as the "correct" rule for this game. Nor do we preclude the possibility that other IP decision rules may offer sensible alternative analyses that deliver a unique strategy at the end, and to help the players avoid the epistemic entanglement. In fact, one may question the merits of the lexicographic rule, on the grounds of information value. It is understood that the lexicographic rule does not necessarily respect the value of cost-free, new information [25]. The mere suggestion by Sidney that they might play the Concert game, despite being a hypothetical one, is enough to steer the game towards the unique outcome that maximizes Sidney's utility globally, but not Isaac's. The answers to some questions remain open for further research. For example: (i) What was Isaac's assessment on the *net value* [15] of his total evidence? and (ii) In general, what should a player do when their total evidence incurs a negative net value?

6. Discussion

We have discussed how imprecise probabilities can help an agent to update their temporal credence with respect to the total evidence, in case a sufficient and observable reduction to it can be found. A question that could have been asked in the first place is whether the sufficiency requirement is necessary. In other words, instead of worrying about finding a sufficient observable event F to serve as a reduction, why don't we consider IP models for the total evidence pair $(F, K_{t,M}(F))$ as in (1)?

The kind of imprecise probabilities that we employ in this paper may not be able to capture *all* varieties of uncertainty and ignorance that a rational agent may have. An IP model is a collection of probabilities defined on a common state space associated with a common sigma field. IP models are useful when the agent is unable to pinpoint their credence function in relation to their corpus of knowledge, nevertheless seeks to update their credence as new information is learned. However, the agent must be certain about their corpus of knowledge, for it is the basis on which to derive any credence at all, precise or otherwise. Expression of uncertainty that pertains to the act of knowing, such as captured by the phrase "I'm not sure if I know F", calls for constructs such as *probabilities of higher types* as advocated by Jack Good [11]. In contrast to IP models, however, higher types of probabilities are fuzzy not only in themselves, but also in the inequalities that can express them. They pose a different challenge in terms of their operationalization, and are therefore out of scope for this paper.

Another open question is whether it is always possible for the agent to find an observable sufficient reduction to their total evidence. We surmise the answer may not be categorically affirmative. In our formulation of the agent's corpus of knowledge (1), the event $K_{t,M}(F)$ that signifies the attainment of the observational report F depend not only on the time of observation t, but also the method M through which the observation of F can be made. For certain method M, to ascertain the event $K_{t,M}(F)$, or any sufficient reduction of it, may well be infeasible for the agent. For example, the measurement of certain complex scientific phenomena is viable only in theory, or may be too costly to perform. Nevertheless, if the agent can identify an affordable and practically observable sufficient reduction for which only ambiguous credence is available, they should prefer it to an unattainable one for which precise credence is available. As demonstrated in this paper, the agent may avoid the epistemic entanglement and extract meaningful inference from the former, using the tools of imprecise probabilities.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank five anonymous reviewers for their helpful comments on the current and earlier versions of this paper, as well as the participants at ISIPTA 2021 for their feedback. The research of R. Gong is supported in part by the National Science Foundation (DMS-1916002).

References

[1] George A. Barnard, The meaning of a significance level, Biometrika 34 (1/2) (1947) 179–182.
[2] Rudolf Carnap, The aim of inductive logic, in: Ernest Nagel, Patrick Suppes, Alfred Tarski (Eds.), Logic, Methodology and Philosophy of Science, Stanford University Press, Stanford, CA, 1962, pp. 303–318, Chapter 5.
[3] Harald Cramér, Mathematical Methods of Statistics (PMS-9), vol. 9, Princeton University Press, 2016.
[4] Persi Diaconis, Review of "A mathematical theory of evidence" (G. Shafer), J. Am. Stat. Assoc. 73 (363) (1978) 677–678.
[5] Persi Diaconis, Sandy L. Zabell, Some alternatives to Bayes's rule, in: Information Pooling and Group Decision Making, JAI Press, 1986, pp. 25–38.
[6] James M. Dickey, Multiple hypergeometric functions: probabilistic interpretations and statistical uses, J. Am. Stat. Assoc. 78 (383) (1983) 628–637.
[7] James M. Dickey, Jhy-Ming Jiang, Joseph B. Kadane, Bayesian methods for censored categorical data, J. Am. Stat. Assoc. 82 (399) (1987) 773–781.
[8] Itzhak Gilboa, David Schmeidler, Maxmin expected utility with non-unique prior, J. Math. Econ. 18 (2) (1989) 141–153.
[9] Raphaël Giraud, Objective imprecise probabilistic information, second order beliefs and ambiguity aversion: an axiomatization, in: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, ISIPTA'05, 2005, pp. 183–192.

[10] Ruobin Gong, Xiao-Li Meng, Judicious judgment meets unsettling updating: dilation, sure loss, and Simpson's paradox (with discussion), Stat. Sci. 36 (2) (2021) 169–190.
[11] Irving John Good, Symposium on current views of subjective probability: subjective probability as the measure of a non-measurable set, in: Studies in Logic and the Foundations of Mathematics, vol. 44, Elsevier, 1966, pp. 319–329.
[12] John C. Harsanyi, Reinhard Selten, A General Theory of Equilibrium Selection in Games, MIT Press Books, vol. 1, 1988.
[13] Howard Hogan, The 1990 post-enumeration survey: operations and results, J. Am. Stat. Assoc. 88 (423) (1993) 1047–1060.
[14] Thomas J. Jiang, Joseph B. Kadane, James M. Dickey, Computation of Carlson's multiple hypergeometric function R for Bayesian applications, J. Comput. Graph. Stat. 1 (3) (1992) 231–251.
[15] Joseph B. Kadane, Mark Schervish, Teddy Seidenfeld, Is ignorance bliss?, J. Philos. 105 (1) (2008) 5–36.
[16] Elon Kohlberg, Jean-François Mertens, On the strategic stability of equilibria, Econometrica 54 (5) (1986) 1003–1037.
[17] Isaac Levi, On indeterminate probabilities, J. Philos. 71 (13) (1974) 391–418.
[18] Isaac Levi, The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance, MIT Press, 1980.
[19] Roderick J.A. Little, Donald B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, 2019.
[20] Hailin Liu, Common knowledge, ambiguity, and the value of information in games, in: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA'15, Pescara, 2015, pp. 167–176.
[21] Hailin Liu, Wei Xiong, Dynamic consistency in incomplete information games under ambiguity, Int. J. Approx. Reason. 76 (2016) 63–79.
[22] Charles F. Manski, Partial Identification of Probability Distributions, Springer Science & Business Media, 2003.
[23] Louis T. Mariano, Joseph B. Kadane, The effect of intensity of effort to reach survey respondents: a Toronto smoking survey, Surv. Methodol. 27 (2) (2001) 131.
[24] Balgobin Nandram, Jai Won Choi, Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability, J. Am. Stat. Assoc. 97 (458) (2002) 381–388.
[25] Teddy Seidenfeld, A contrast between two decision rules for use with (convex) sets of probabilities: Γ-maximin versus E-admissibility, Synthese 140 (1/2) (2004) 69–88.
[26] Teddy Seidenfeld, Larry Wasserman, Dilation for sets of probabilities, Ann. Stat. 21 (3) (1993) 1139–1154.
[27] Steve Selvin, On the Monty Hall problem (letter to the editor), Am. Stat. 29 (3) (1975) 134.
[28] Steve Selvin, A problem in probability (letter to the editor), Am. Stat. 29 (1) (1975) 67.
[29] Elizabeth A. Stasny, Hierarchical models for the probabilities of a survey classification and nonresponse: an example from the National Crime Survey, J. Am. Stat. Assoc. 86 (414) (1991) 296–303.
[30] Matthias C.M. Troffaes, Decision making under uncertainty using imprecise probabilities, Int. J. Approx. Reason. 45 (1) (2007) 17–29.
[31] Marilyn vos Savant, Game show problem, https://web.archive.org/web/20120429013941/http://marilynvossavant.com/game-show-problem/. (Accessed 14 February 2021).