ARTICLE IN PRESS

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Safe probability

Peter Grünwald *

CWI, Science Park 123, 1098 XG, Amsterdam, The Netherlands
The Netherlands and Leiden University, Mathematical Institute Niels, Bohrweg 1, 2333 CA Leiden, The Netherlands

ARTICLE INFO

Article history:
Available online xxxx

ABSTRACT

We formalize the idea of probability distributions that lead to reliable predictions about some, but not all aspects of a domain. The resulting notion of 'safety' provides a fresh perspective on foundational issues in statistics, providing a middle ground between imprecise probability and multiple-prior models on the one hand and strictly Bayesian approaches on the other. It also allows us to formalize fiducial distributions in terms of the set of random variables that they can safely predict, thus taking some of the sting out of the fiducial idea. By restricting probabilistic inference to safe uses, one also automatically avoids paradoxes such as the Monty Hall problem. Safety comes in a variety of degrees, such as 'validity' (the strongest notion), 'calibration', 'confidence safety' and 'unbiasedness' (almost the weakest notion).

1. Introduction

We formalize the idea of probability distributions that lead to reliable predictions about some, but not all aspects of a domain. Very broadly speaking, we call a distribution $\tilde{P}$ *safe* for predicting random variable U given random variable V if predictions concerning U based on $\tilde{P}(U|V)$ tend to be as good as one would expect them to be if $\tilde{P}$ were an accurate description of one's uncertainty, even if $\tilde{P}$ may not represent one's actual beliefs, let alone the truth. Our formalization of this notion of 'safety' has repercussions for the foundations of statistics, providing a joint perspective on issues hitherto viewed as distinct:

1. All models are wrong...[1] Some statistical models are evidently both entirely wrong yet very useful. For example, in some highly successful applications of Bayesian statistics, such as latent Dirichlet allocation for topic modeling (Blei et al., 2003), one assumes that natural language text is i.i.d., which is fine for the task at hand (topic modeling) — yet no-one would want to use these models for predicting the next word of a text given the past. Yet, one can use a Bayesian posterior to make such predictions any way — Bayesian inference has no mechanism to distinguish between 'safe' and 'unsafe' inferences. Safe probability allows us to impose such a distinction.

2. The eternal discussion[2] More generally, representing uncertainty by a single distribution, as is standard in Bayesian inference, implies a willingness to make definite predictions about random variables that, some claim, one really knows nothing about. Disagreement on this issue goes back at least to Keynes (1921) and Ramsey (1931), has led many economists to sympathize with *multiple-prior models* (Gilboa and Schmeidler, 1989) and some statisticians to embrace the related

* Correspondence to: CWI, Science Park 123, 1098 XG, Amsterdam, United Kingdom.
E-mail address: pdg@cwi.nl.

1 ...yet some are useful, as famously remarked by Box (1979).

2 When the single- vs. multiple-prior issue came up in a discussion on the *decision-theory forum* mailing list, the well-known economist I. Gilboa referred to it as 'the eternal discussion'.

https://doi.org/10.1016/j.jspi.2017.09.014
0378-3758/

imprecise probability (Walley, 1991; Augustin et al., 2014) in which so-called 'Knightian' uncertainty is modeled by a *set* $\mathcal{P}^*$ of distributions. But imprecise probability is not without problems of its own, an important one being *dilation* (Example 1). Safe probability can be understood as starting from a set $\mathcal{P}^*$, but then *mapping* the set of distributions to a single distribution, where the mapping invoked may depend on the prediction task at hand — thus avoiding both dilation and overly precise predictions. The use of such mappings has been advocated before, under the name *pignistic transformation* (Smets, 1989; Hampel, 2001), but a general theory for constructing and evaluating them has been lacking (see also Section 4).

3. Fisher's Biggest Blunder[3] Fisher (1930) introduced *fiducial inference*, a method to come up with a 'posterior' $\tilde{P}(\theta \mid X^n)$ on a model's parameter space based on data X^n, but without anything like a 'prior', in an approach to statistics that was neither Bayesian nor frequentist. The approach turned out problematic however, and, despite progress on related *structural inference* (Fraser, 1968, 1979) it was largely abandoned. Recently, however, fiducial distributions have made a comeback (Hannig, 2009; Taraldsen and Lindqvist, 2013; Martin and Liu, 2013; Veronese and Melilli, 2015), in some instances with a more modest, frequentist interpretation as *confidence distributions* (Schweder and Hjort, 2002, 2016). As noted by Xie and Singh (2013), these 'contain a wealth of information for inference', e.g. to determine valid confidence intervals and unbiased estimation of the median, but their interpretation remains difficult, viz. the insistence by Hampel (2006) and Xie and Singh (2013) and many others that, although $\tilde{P}(\cdot \mid X^n)$ is defined as a distribution on the parameter space, the parameter itself is not random. Safe probability offers an alternative perspective, where the insistence that 'θ is not random' is replaced by the weaker (and perhaps liberating) statement that 'we can treat θ as random' *as long as we restrict ourselves to safe inferences about it* — in Section 3.1 we determine precisely what these safe inferences are and how they fit into a general hierarchy:

4. The Hierarchy Pursuing the idea that some distributions are reliable for a smaller subset of random variables/prediction tasks than others, leads to a natural *hierarchy* of safeties — a first taste of which is in Fig. 1, with notations explained later. At the top are distributions that are fully reliable for whatever task one has in mind; at the bottom those that are reliable only for a single task in a weak, average sense. In between there is a natural place for distributions that are *calibrated* (Example 2), that are *confidence-safe* (i.e. valid confidence distributions) and that are *optimal for squared-error prediction*.

5. "The concept of a conditional probability with regard to an isolated hypothesis..."[4] Upon first hearing of the Monty Hall (quiz master, three doors) problem (Vos Savant, 1990; Gill, 2011), most people naively think that the probability of winning the car is the same whether one switches doors or not. Most can eventually, after much arguing, be convinced that this is wrong, but wouldn't it be nice to have a simple sanity check that *immediately* tells you that the naive answer must be wrong, without even pondering the 'right' way to approach the problem? Safe probability provides such a check: one can immediately tell that the naive answer is *not safe*, and thus cannot be right. Such a check is applicable more generally, whenever conditioning on events rather than on random variables. This safety check is based on the developments in this paper, but, to keep the length of this paper at bay, we will fully report on it elsewhere — and will only briefly return to it in the concluding Section 4.

6. Further Applications: Objective Bayes, Epistemic Probability, Hypothesis Testing Apart from the applications above, the results in this paper suggest that safe probability be used to formalize the status of default priors in 'objective Bayesian' inferences, and to enable an alternative look at *epistemic probability*. There are also strong repercussions for *hypothesis testing* (Is the conclusion still 'safe' under optional stopping?). These are both topics for future work, and we briefly return to them in Section 4. Finally, we mention that, although all applications in this paper can be understood as using $\tilde{P}$'s which are in some sense misspecified, we do not consider in this paper the 'standard' misspecification case in which one bases statistical inference on a model that is wrong-yet-useful, such as the Dirichlet model in topic modeling mentioned above. Safe probability is eminently suited to model this type of application as well though — this has already partially been done (Grünwald and van Ommen, 2014), and again, we briefly return to this point in Section 4.

Starting with Grünwald (1999), my own work – often in collaboration with J. Halpern – has regularly used the idea of 'safety', for example in the context of Maximum Entropy inference (Grünwald, 2000), and also dilation (Grünwald and Halpern, 2004), calibration (Grünwald and Halpern, 2011), and probability puzzles like Monty Hall (Grünwald and Halpern, 2003; Grünwald, 2013). However, the insights of earlier papers were very partial and scattered, and the present paper presents for the first time a general formalism, definitions and a hierarchy. It is also the first one to make a connection to confidence distributions and pivots. Further applications, as indicated above, are mostly postponed to future papers.

1.1. Informal overview

Below we explain the basic ideas using three recurring examples. We assume that we are given a set of distributions $\mathcal{P}^*$ on some space of outcomes $\mathcal{Z}$. Under a frequentist interpretation, $\mathcal{P}^*$ is the set of distributions that we regard as 'potentially

[3] While Fisher is generally regarded as (one of) the greatest statisticians of all time, fiducial inference is often considered to be his 'big blunder' — see Hampel (2006) and Efron (1996), who writes *Maybe Fisher's biggest blunder will become a big hit in the 21st century!*.

[4] ... whose probability equals 0 is inadmissible", as remarked by Kolmogorov (1933). Safe probability suggests that there are in fact more cases in which it is inadmissible, and these are related to the Monty Hall sanity check.

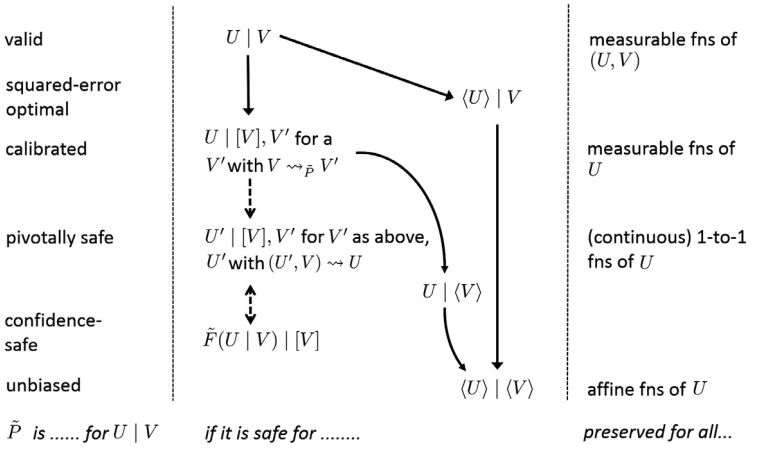

Fig. 1. A Hierarchy of Relations for $\tilde{P}$ — notation and formalism defined and explained in later sections. The concepts on the right correspond (broadly) to existing notions, whose name is given on the left (with the exception of $U \mid \langle V \rangle$, for which no regular name seems to exist). $A \rightarrow B$ means that safety of $\tilde{P}$ for A implies safety for B — at least, under some conditions: for all solid arrows, this is proven under the assumption of V with countable range (see underneath Proposition 1). For the dashed arrows, this is proven under additional conditions (see Theorem 2 and subsequent remark). On the right are shown transformations on U under which safety is preserved, e.g. if $\tilde{P}$ is calibrated for $U|V$ then it is also calibrated for $U' \mid V$ for every U' with $U \rightsquigarrow U'$ (see remark underneath Theorem 2). Weakening the conditions for the proofs and providing more detailed interrelations is a major goal for future work, as well as investigating whether the hierarchy has a natural place for *causal* notions, such as $\tilde{P}(U \mid \mathrm{DO}(v))$ as in Pearl's (2009) do-calculus.

true'; under a subjectivist interpretation, it is the *credal set* that describes our uncertainty or 'beliefs'; all developments below work under both interpretations.

All probability distributions mentioned below are either an element of $\mathcal{P}^*$, or they are a *pragmatic distribution* $\tilde{P}$, which some decision-maker (DM) uses to predict the outcomes of some variable U given the value of some other variable V,where both U and V are random quantities defined on $\mathcal{Z}$. $\tilde{P}$ is also used to estimate the quality of such predictions. $\tilde{P}$ (which may be, but is not always in $\mathcal{P}^*$) is 'pragmatic' because we assume from the outset that some element of $\mathcal{P}^*$ might actually lead to better predictions — we just do not know which one.

Example 1 (*Dilation*)**.** A DM has to make a prediction or decision about random variable $U \in \mathcal{U} = \{0, 1\}$ given the value of $V \in \mathcal{V} = \{0, 1\}$. She knows that the marginal probability $P(U = 1) = 0.9$; she suspects that U may depend on V, but has no idea whether U and V are positively or negatively correlated or how strong the correlation is. She may thus model her uncertainty as the set $\mathcal{P}^*$ of *all* distributions P on $\mathcal{Z} = \mathcal{U} \times \mathcal{V}$ that satisfy

$$P(U = 1) = \sum_{v \in \mathcal{V}} P(U = 1, V = v) = 0.9. \tag{1}$$

Given that $V = 1$, what should she predict for U? A standard answer in imprecise probability (Walley, 1991) is to pointwise condition the set $\mathcal{P}^*$, leading one to adopt the probabilities $\mathcal{P}^*(U = 1 \mid V = 1) := \{P(U = 1 \mid V = 1) : P \in \mathcal{P}^*\}$. But this set contains *every* distribution on U, including $P(U = 1 \mid V = 1) = 0$ (the latter would be obtained for the $P \in \mathcal{P}^*$ with $P(U = |1 - V|) = 1$). It therefore seems that, after observing $V = 1$, the DM has lost rather than gained information. By symmetry, the same happens after observing $V = 0$, so *whatever DM observes, she loses information* — a phenomenon known as *dilation* (Seidenfeld and Wasserman, 1993). This is intuitively disturbing, and it may perhaps be better to simply *ignore* V and predict using the distribution that acts as if $U \perp V$ and has

$$\tilde{P}(U = 1 \mid V = v) = P(U = 1) \text{ for all } v \in \mathcal{V}, \tag{2}$$

i.e. $\tilde{P}(U = 1 \mid V = v) = 0.9$. While from a purely subjective Bayesian standpoint information is never useless and this seems silly, it is certainly what humans often do in practice, and usually, they get away with it (Dempster, 1968) — for concrete examples see Grünwald and Halpern (2004). Here is where Safe Probability comes in — it tells us that $\tilde{P}$ is *safe* to use, in the following simple sense: for any function $g : \mathcal{U} \rightarrow \mathbb{R}$, we have:

$$\text{for all } P \in \mathcal{P}^*, \text{ all } v \in \mathcal{V}: \quad \mathrm{E}_{U \sim P}[g(U)] = \mathrm{E}_{U \sim \tilde{P}}[g(U) \mid V = v]. \tag{3}$$

In particular, if we have a loss function $L : \mathcal{U} \times \mathcal{A} \rightarrow \mathbb{R}$ mapping outcomes and actions to associated losses, then, for any action $a \in \mathcal{A}$, we can plug in $g(U) := L(U, a)$ above and then we find that (assuming $\mathcal{P}^*$ contains the truth):

> DM's predictions are guaranteed to be exactly as good, in expectation, as *she would expect them to be if* $\tilde{P}$ *were actually 'true'* — even if $\tilde{P}$ is not true at all.

We immediately add though that if we had a loss function $L' : \mathcal{U} \times \mathcal{V} \times \mathcal{A} \to \mathbb{R}$ which would *itself* depend on V (e.g. if $V = 1$ DM is offered a different bet on U than if $V = 0$) then the $\tilde{P}$ based on ignoring V is not safe any more — (3) may not hold any more, and the actual expectation may be different from DM's. In terms of the formalism we develop below (Definitions 1, 2 and 3), this will be expressed as '$\tilde{P}$ is safe for predicting with loss function L but not loss function L'', or, in formal notation, $\tilde{P}$ is safe for $L(\cdot, a) \mid [V]$ but not for $L'(\cdot, a) \mid [V]$. The intuitive meaning is that DM can safely use $\tilde{P}$ to make predictions against L (her predictions will be as good as she expects) but not against L'. These statements will be immediate consequences of the more general statements "$\tilde{P}$ is safe for $U \mid [V]$ but not safe for $U \mid V$".

In some cases, we will not be able to come up with a $\tilde{P}$ satisfying (3), and we have to settle for a $\tilde{P}$ that satisfies a weaker notion of safety, such as, for all $P \in \mathcal{P}^*$, all functions g,

$$\mathrm{E}_{V \sim P}\left[\mathrm{E}_{U \sim \tilde{P}}[g(U) \mid V]\right] = E_{U \sim P}[g(U)], \tag{4}$$

which says that DM predicts as well on average as DM would expect to predict on average if $\tilde{P}$ were true, even though $\tilde{P}$ may not be true. This will be denoted as '$\tilde{P}$ is safe for $U \mid \langle V \rangle$'; and if (4) only holds for g the identity (which makes no difference if $|\mathcal{U}| = 2$, but in general it does) we have the even weaker safety for $\langle U \rangle \mid \langle V \rangle$ (Fig. 1). In Section 2.2 we thus obtain five basic notions of safety, varying from weak safety, in an average sense, to very strong safety, safety for $U \mid V$, which essentially means that $\tilde{P}(U \mid V)$ must be the correct conditional distribution.

In this example we used frequentist terminology, such as 'correct' and 'true', and we continue to do so in this paper. Still, a subjective interpretation remains valid in this and future examples as well: if the DM's real beliefs are given by the full set $\mathcal{P}^*$, she can safely act as if her belief is represented by the singleton $\tilde{P}$ as long as she also believes that her loss does not depend on V.

Example 2 (*Calibration*)**.** Consider the weather forecaster on your local television station. Every night the forecaster makes a prediction about whether or not it will rain the next day in the area where you live. She does this by asserting that the probability of rain is p, where $p \in \{0, 0.1, \ldots, 0.9, 1\}$. How should we interpret these probabilities? The usual interpretation is that, in the long run, on those days at which the weather forecaster predicts probability p, it will rain approximately $100p\%$ of the time. Thus, for example, among all days for which she predicted 0.1, the fraction of days with rain was close to 0.1. A weather forecaster (DM) with this property is said to be *calibrated* (Dawid, 1982; Foster and Vohra, 1998). Like safety itself, calibration is a *minimal* requirement: for example, a weather forecaster who predicts, each day of the year, that the probability of rain tomorrow is 50% will be approximately calibrated in the Netherlands, but her predictions are not very useful — and it is easily seen that, when using a proper scoring rule, optimal forecasts are calibrated, but calibrated forecasts can be far from optimal. On the other hand, in practice we often see calibrated weather forecasters that predict well, but do not predict with anything close to the 'truth' — their predictions depend on high-dimensional covariates consisting of measurements of air pressure, temperature etc. at numerous locations in the world, and it seems quite unlikely (and, for practical purposes, unnecessary!) that, given any specific values of these covariates, they issue the correct conditional distribution. While calibration is usually defined relative to empirical data, a re-definition in terms of an underlying set of distributions $\mathcal{P}^*$ is straightforward (Vovk et al., 2005; Grünwald and Halpern, 2011), and in Section 2.3 we show that the probabilistic definition of calibration has a natural expression in terms of the safety notions introduced above: $\tilde{P}(U \mid V)$ is calibrated for U if it is safe for $U \mid [V]$, V', for *some* V' with $V \rightsquigarrow V'$ (all notation to be explained) — which implies that (3) is itself an instance of calibration.

Example 3 (*Bayesian, Fiducial and Confidence Distributions*)**.** We are given a parametric probability model $\mathcal{M} = \{q_\theta \mid \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$, each q_θ defines a probability density or mass function on data $(X_1, \ldots, X_N) = X^N$ of sample size N, each outcome X_i taking a value in some space $\mathcal{X}$. The goal is to make inferences about θ, based on the data X^N or some statistic $S(X^N, N)$ thereof. In the common case with fixed $N = n$ and inference based on the full data, $S(X^N, N) := X^n$, we can transfer this statistical scenario to our setup by defining $\mathcal{P}^*$ as a set of distributions on $\mathcal{Z} = \Theta \times \mathcal{X}^n$. RVs U and $V = S(X^n, n) = X^n$ are then defined as, for each $z = (\theta, x^n)$, $U(z) := \theta$ and $V(z) := x^n$. DM employs a set Π of prior distributions on Θ, where each $\pi \in \Pi$ induces a joint distribution P_π on $\Theta \times \mathcal{X}^n$ with marginal on Θ determined by π and, given θ, density of x^n given by q_θ, so that if π has density p_π, we get the joint density $p_\pi(\theta, x^n) = p_\pi(\theta) \cdot q_\theta(x^n)$. We set $\mathcal{P}^* := \{P_\pi : \pi \in \Pi\}$ to be the set of all such joint distributions. In the special case in which DM really is a 100% subjective Bayesian who believes that a single prior π captures all uncertainty, we have that $\mathcal{P}^* = \{P_\pi\}$ contains just a single joint parameter-data distribution, and we are in the standard Bayesian scenario. Then DM can set $\tilde{P}(\theta \mid X^n) := P_\pi(\theta \mid X^n)$, the standard posterior, and any type of inference about θ is safe relative to $\mathcal{P}^*$. Here we focus on another special case, in which Π contains exactly one density for each $\theta \in \Theta$, namely the degenerate distribution putting all its mass on θ. We denote this distribution by P_θ and notice that then $\mathcal{P}^* = \{P_\theta : \theta \in \Theta\}$, with $P_\theta(\Theta = \theta) = 1$, and for any measurable set $\mathcal{A}$, $P_\theta(X^n \in \mathcal{A})$ determined by density p_θ, satisfying

$$p_\theta(x^n) = p_\theta(x^n \mid \Theta = \theta) = q_\theta(x^n).$$

Still, any choice of pragmatic distribution $\tilde{P}(U \mid V) = \tilde{P}(\theta \mid X^n)$ can be interpreted as a distribution on $U \equiv \Theta$ given the data X^n, analogous to a Bayesian posterior (another option, to which we return in Section 4, is to view the model as potentially misspecified, and choose $\mathcal{P}^*$ to be strictly larger than $\mathcal{M}$). In Section 3 we investigate how one can construct distributions $\tilde{P}$

of this kind (only for the well-specified case!) that are safe for inference about *confidence intervals*. for simplicity we restrict ourselves to the 1-dimensional case, for which we find that the construction we provide leads to $\tilde{P}$ that are confidence-safe, written in our notation as 'safe for $\tilde{F}(U|V) \mid [V]$', with $\tilde{F}$ being the CDF (cumulative distribution function) of $\tilde{P}(U|V)$. Confidence safety is roughly the same as coverage (Sweeting, 2001): it means that the 'true' probability that θ is contained in a particular type of α-credible sets (sets with 'posterior' probability α given the data V), is equal to α.

The $\tilde{P}$ we construct are essentially equivalent to the *confidence distributions* of Schweder and Hjort (2002), that were designed with the explicit goal of having good confidence properties; they also often coincide with Fisher's (1930) fiducial distributions, which in later work (Fisher, 1935) he started treating as ordinary probability distributions that could be used without any restrictions. This cannot be right (see e.g. (Hampel, 2006, page 514)), but the question has always remained how a probability calculus for fiducial distributions could be derived that incorporates the right restrictions. Our work provides a step in this direction, in that we show how such $\tilde{P}$ snugly fit into our general framework: confidence safety is a strictly weaker property than calibration, and has again a natural representation in terms of the $\langle U \rangle \mid \langle V \rangle$ notation mentioned above. Moreover, it is a special case of *pivotal safety* which also has repercussions in quite different contexts such as the Monty Hall problem, to which we return in Section 4.

The example illustrates two important points:

1. In some cases the literature suggests some method for constructing a pragmatic $\tilde{P}$. An example is the latent Dirichlet allocation model (Blei et al., 2003) mentioned above, in which data V are text corpora, $\mathcal{P}^*$, not explicitly given, is a complicated set of realistic distributions over V under which data are non-i.i.d., and the literature suggests to take $\tilde{P}(U \mid V)$ as the Bayesian posterior for a cleverly designed i.i.d. model.
2. In other cases, DM may want to construct a $\tilde{P}$ herself. In Example 1, the safe $\tilde{P}$ was obtained by replacing an (unknown) conditional distribution with a (known) marginal — a special case of what was called $\mathcal{C}$-conditioning by Grünwald and Halpern (2011). Marginal distributions and distributions that ignore aspects of V play a more central role in this construction process: they also do in the confidence construction mentioned above, where one sets $\tilde{P}(U \mid V)$ equal to a distribution such that $\tilde{P}(U' \mid V)$, where U' is some auxiliary random variable (a *pivot*), becomes independent of V. For the original RV U though, in the dilation example, DM acts as if U and V are independent even though they may not be; in the confidence distribution example, DM acts in a 'dual' manner, namely as if U and V are dependent, even though under $\mathcal{P}^*$ they are not — which is fine, as long as her conclusions are *safe*.

Overview of the paper In Section 2, we treat the case of countable space $\mathcal{Z}$, defining the basic notions of safety in Section 2.2 (where we return to dilation), and showing how calibration can be cleanly expressed using our notions in Section 2.3. In Section 3 we extend the setting to general $\mathcal{Z}$, which is needed to handle the case of confidence safety (Section 3.1), pivots (Section 3.2) and squared error optimality, where we observe continuous-valued random variables. We end with a discussion of further potential applications of safety as well as open problems. Proofs and further technical details are delegated to Appendix A.

2. Basic definitions for discrete random variables

For simplicity, we introduce our basic notions only considering countable $\mathcal{Z}$, which allows us to sidestep measurability issues altogether. Thus below, $\mathcal{Z}$ is countable; we treat the general case in Section 3.

2.1. Concepts and notations regarding distributions on $\mathcal{Z}$

We define a random variable (abbreviated to RV) to be any function $X : \mathcal{Z} \to \mathbb{R}^k$ for some $k > 0$. Thus RVs can be multidimensional (i.e. what is usually called 'random vector'). By an '$\mathcal{Y}$-valued RV' or simply 'generalized RV' we mean any function mapping $\mathcal{Z}$ to an arbitrary set $\mathcal{Y}$. For two RVs $U = (U_1, \ldots, U_{k_1}), V = (V_1, \ldots, V_{k_2})$ where U_j and V_j are 1-dimensional random variables, we define (U, V) to be the RV with components $(U_1, \ldots, U_{k_1}, V_1, \ldots, V_{k_2})$.

For any generalized RVs U and V on $\mathcal{Z}$ and function f we write $U \overset{f}{\rightsquigarrow} V$ if for all $z \in \mathcal{Z}$, $V(z) = f(U(z))$. We write $U \rightsquigarrow V$ ("U determines V", or equivalently "U is a *coarsening* of V") if there is a function f such that $U \overset{f}{\rightsquigarrow} V$. We write $U \leftrightsquigarrow V$ if $U \rightsquigarrow V$ and $V \rightsquigarrow U$. For two GRVs U and V we write $U \equiv V$ if they define the same function on $\mathcal{Z}$, and for a distribution $P \in \mathcal{Z}$ we write $U =_P V$ if $P(U = V) = 1$. We write $U \overset{f}{\rightsquigarrow}_P V$ if $P(\{z \in \mathcal{Z} : V(z) = f(U(z))\}) = 1$, and $U \rightsquigarrow_P V$ if there exists some f for which this holds. Clearly $U \rightsquigarrow V$ implies that for all distributions P on $\mathcal{Z}$, $U \rightsquigarrow_P V$, but not vice versa. Let $S : \mathcal{Z} \to \mathcal{S}$ be a function on $\mathcal{Z}$. The *range* of S, denoted $\textsc{range}(S)$, the *support* of S under a distribution P, and the range of S given that another function T on $\mathcal{Z}$ takes value t, are denoted as

$$\textsc{range}(S) := \{s \in \mathcal{S} : s = S(z) \text{ for some } z \in \mathcal{Z}\}; \ \textsc{supp}_P(S) := \{s \in \mathcal{S} : P(S = s) > 0\},$$
$$\textsc{range}(S \mid T = t) = \{s \in \mathcal{S} : s = S(z) \text{ for some} z \in \mathcal{Z} \text{ with } t = T(z)\} \tag{5}$$

where we note that $\textsc{supp}_P(S) \subseteq \textsc{range}(S)$, with equality if S has full support.

For a distribution P on $\mathcal{Z}$, and $\mathcal{U}$-valued RV U, we write $P(U)$ as short-hand to denote the distribution of U under P (i.e. $P(U)$ is a probability measure).

We generally omit double brackets, i.e. if we write $P(U, W)$ for RVs U and W, we really mean $P(R)$ where R is the RV (U, W),

Any generalized RV that maps all $z \in \mathcal{Z}$ to the same constant is called *trivial*, in particular the RV $\mathbf{0}$ which maps all $z \in \mathcal{Z}$ to 0. For an event $\mathcal{E} \subset \mathcal{Z}$, we define the *indicator random variable* $\mathbf{1}_{\mathcal{E}}$ to be 1 if $\mathcal{E}$ holds and 0 otherwise.

Conditional distributions as generalized RVS For given distribution on $\mathcal{Z}$ and generalized RVs V and W, we denote, for all $v \in \text{SUPP}_P(V)$, $P \mid V = v$ as the conditional distribution on $\mathcal{Z}$ given $V = v$, in the standard manner. We further define $(\mathcal{P}^* \mid W = w) := \{(P \mid W = w) : P \in \mathcal{P}^*, w \in \text{SUPP}_P(W)\}$ to be the set of distributions on $\mathcal{Z}$ that can be arrived at from $\mathcal{P}$ by conditioning on $W = w$, for all w supported by some $P \in \mathcal{Z}$.

We further denote, for all $v \in \text{SUPP}_P(V)$, $P(U \mid V = v)$ as the conditional distribution of U given $V = v$, defined as the distribution on U given by $P(U = u \mid V = v) := P(U = u, V = v)/P(V = v)$ (whereas $P \mid V = v$ is defined as a distribution on $\mathcal{Z}$, $P(U \mid V = v)$ is a distribution on the more restricted space $\text{RANGE}(U)$).

Suppose DM is interested in predicting RV U given RV V and does this using some conditional distribution $P(U \mid V = v)$ (usually this P will be the 'pragmatic' $\tilde{P}$, but the definition that follows holds generally). Adopting the standard convention for conditional expectation, we call any function from $\text{RANGE}(V)$ to the set of distributions on U that coincides with $P(U \mid V = v)$ for all $v \in \text{SUPP}_P(V)$ a *version* of the conditional distribution $P(U \mid V)$. If we make a statement of the form '$P(U \mid V)$ satisfies ...', we really mean 'every version of $P(U \mid V)$ satisfies...'. We thus treat $P(U \mid V)$ as a $\mathcal{E}$-valued random variable where $\mathcal{E} = \{P(U \mid V = v) : v \in \text{RANGE}(V)\}$, where, for all $z \in \mathcal{Z}$ with $P(V = V(z)) > 0$, $P(U \mid V)(z) := P(U \mid V = V(z))$, and $P(U \mid V)(z)$ set to an arbitrary value otherwise.

Unique and well-definedness Recall that DM starts with a set $\mathcal{P}^*$ of distributions on $\mathcal{Z}$ that she considers the right description of her uncertainty. She will predict some RV U given some generalized RV V using a *pragmatic* distribution $\tilde{P}$.

For RV $U : \mathcal{Z} \to \mathbb{R}^k$ and generalized RV V, we say that, for given distribution P' on $\mathcal{Z}$, $P'(U \mid V)$ is *essentially uniquely defined* (relative to $\mathcal{P}^*$) if for all $P \in \mathcal{P}^*$, $\text{SUPP}_P(V) \subseteq \text{SUPP}_{P'}(V)$ (so that P-almost surely V takes value v with $P'(V = v) > 0$). We use this definition both for $P' \in \mathcal{P}^*$ and for $P' = \tilde{P}$; note that we always *evaluate* whether P' is uniquely defined under distributions in the 'true' $\mathcal{P}^*$ though.

We say that $\text{E}_{P'}[U \mid V]$ is well-defined if, writing $U = (U_1, \ldots, U_k)$, and, $U_j^+ = \max\{U_j, 0\}$, $U_j^- = \max\{-U_j, 0\}$, we have, for $j = 1..k$, either $\text{E}_{P'}[U_j^+ \mid V] < \infty$ with P-probability 1, or $\text{E}_{P'}[U_j^- \mid V] < \infty$ with P-probability 1. This is a very weak requirement that ensures that calculating expectations never involves the operation $\infty - \infty$, making all expectations well-defined.

The pragmatic distribution $\tilde{P}$ We assume that DM makes her predictions based on a probability distribution $\tilde{P}$ on $\mathcal{Z}$ which we generally refer to as the *pragmatic distribution*. In practice, DM will usually be presented with a decision problem in which she has to predict some fixed RV U based on some fixed RV V, and then she is only interested in the conditional distribution $\tilde{P}(U \mid V)$, and for some other RVs U' and V', $\tilde{P}(U' \mid V')$ may be left undefined. In other cases she only may want to predict the expectation of U given V — in that case she only needs to specify $\text{E}_{\tilde{P}}[U \mid V]$ as a function of V, and all other details of $\tilde{P}$ may be left unspecified. In Appendix A.1 we explain how to deal with such *partially specified* $\tilde{P}$. In the main text though, for simplicity we assume that $\tilde{P}$ is a fully-specified distribution on $\mathcal{Z}$; DM can fill up irrelevant details any way she likes. The very goal of our paper being to restrict $\tilde{P}$ to making 'safe' predictions however, DM may come up with $\tilde{P}$ to predict U given V and there may be many RVs U' and V' definable on the domain such that $\tilde{P}(U' \mid V')$ has no bearing to $\mathcal{P}^*$ and would lead to terrible predictions; as long as we make sure that $\tilde{P}$ is not used for such U' and V' – which we will – this will not harm the DM.

2.2. The basic notions of safety

All our subsequent notions of 'safety' will be constructed in terms of the following first, simple definitions.

Definition 1 (*Weak Safety Notions*)**.** Let $\mathcal{Z}$ be an outcome space and $\mathcal{P}^*$ be a set of distributions on $\mathcal{Z}$, let U be an RV and V be a generalized RV on $\mathcal{Z}$, and let $\tilde{P}$ be a distribution on $\mathcal{Z}$. We say that $\tilde{P}$ is *safe* for $\langle U \rangle \mid [\![V]\!]$ (pronounced as '$\tilde{P}$ is safe for *predicting* $\langle U \rangle$ *given* $[\![V]\!]$'), if

$$\text{for all } P \in \mathcal{P}^* : \inf_{v \in \text{SUPP}_{\tilde{P}}(V)} \text{E}_{\tilde{P}}[U|V = v] \leq \text{E}_P[U] \leq \sup_{v \in \text{SUPP}_{\tilde{P}}(V)} \text{E}_{\tilde{P}}[U|V = v]. \tag{6}$$

We say that $\tilde{P}$ is *safe* for $\langle U \rangle \mid \langle V \rangle$, if

$$\text{for all } P \in \mathcal{P}^* : \text{E}_P[U] = \text{E}_P[\text{E}_{\tilde{P}}[U|V]]. \tag{7}$$

We say that $\tilde{P}$ is *safe* for $\langle U \rangle \mid [V]$, if (6) holds with both inequalities replaced by an equality, i.e. for all $v \in \text{SUPP}_{\tilde{P}}(V)$,

$$\text{for all } P \in \mathcal{P}^* : \text{E}_P[U] = \text{E}_{\tilde{P}}[U|V = v]. \tag{8}$$

In this definition, as in all definitions and results to come, whenever we write '⟨ statement ⟩' we really mean 'all conditional probabilities in the following statement are essentially uniquely defined, all expectations are well-defined, and ⟨ statement ⟩'. Hence, (7) really means 'for all $P \in \mathcal{P}^*$, $\tilde{P}(U|V)$ is essentially uniquely defined, $\mathrm{E}_{\tilde{P}}[U|V]$, $\mathrm{E}_P[U]$, and $\mathrm{E}_P[\mathrm{E}_{\tilde{P}}[U|V]]$ are well-defined, and the latter two are equal to each other'. Also, when we wrote $\tilde{P}$ is safe for $\langle U\rangle \mid \langle V\rangle$, we really meant that it is safe for $\langle U\rangle \mid \langle V\rangle$ *relative to the given* $\mathcal{P}^*$; we will in general leave out the phrase 'relative to $\mathcal{P}^*$', whenever this cannot cause confusion.

To be fully clear about notation, note that in double expectations like in (7), we consider the right random variable to be bound by the outer expectation; thus it can be rewritten in any of the following ways:

$$\mathrm{E}_{U\sim P}[U] = \mathrm{E}_{V\sim P}\mathrm{E}_{U\sim \tilde{P}|V}[U]$$
$$\mathrm{E}_{V\sim P}\mathrm{E}_{U\sim P|V}[U] = \mathrm{E}_{V\sim P}\mathrm{E}_{U\sim \tilde{P}|V}[U]$$
$$\sum_{u\in \mathrm{RANGE}(U)} P(U=u)\cdot u = \sum_{v\in \mathrm{RANGE}(V)} P(V=v)\cdot \sum_{u\in \mathrm{RANGE}(U)} \tilde{P}(U=u \mid V=v)\cdot u,$$

where the second equality follows from the tower property of conditional expectation.

Towards a hierarchy It is immediately seen that, if $\tilde{P}$ is safe for $\langle U\rangle \mid [V]$, then it is also safe for $\langle U\rangle \mid \langle V\rangle$, and if it is safe for $\langle U\rangle \mid \langle V\rangle$, then it is also safe for $\langle U\rangle \mid [\![V]\!]$. This hierarchy will be extended below Proposition 1 and then gives rise to Fig. 1. Safety for $\langle U\rangle \mid [\![V]\!]$ is thus the weakest notion — it allows a DM to give valid upper- and lower-bounds on the actual expectation of U, by quoting $\sup_{v\in \mathrm{SUPP}_{\tilde{P}}(V)}\mathrm{E}_{\tilde{P}}[U \mid V=v]$ and $\inf_{v\in \mathrm{SUPP}_{\tilde{P}}(V)}\mathrm{E}_{\tilde{P}}[U \mid V=v]$, respectively, but nothing more. It will hardly be used here, except for Example 10; it plays an important role though in applications of safety to hypothesis testing, on which we hope to report in future work.

Safety for $\langle U\rangle \mid \langle V\rangle$ evidently bears relations to *unbiased* estimation: if $\tilde{P}$ is safe for $\langle U\rangle \mid \langle V\rangle$, i.e. (7) holds, then we can think of $\mathrm{E}_{\tilde{P}}[U|V]$ as an unbiased estimate, based on observing V, of the random quantity U (see also Example 7 later on). Safety for $\langle U\rangle \mid [V]$ implies that all distributions in $\mathcal{P}^*$ agree on the expectation of U and that $\mathrm{E}_{\tilde{P}}[U \mid V=v]$ is the same for (essentially) all values of v, and is thus a much stronger notion.

Example 4 (*Dilation: Example 1, Cont.*)**.** The first application of definition (7) was already given in Example 1, where we used a $\tilde{P}$ that ignored V and was safe for $\langle U\rangle \mid \langle V\rangle$ and $\langle U\rangle \mid [V]$, as we see from (4) with g the identity. Let us extend the example, replacing $\mathcal{U}=\{0,1\}$ in that example by $\mathcal{U}=\{0,1,2\}$, with $\mathcal{P}^*$ again defined as the set of all distributions satisfying (1) and $\tilde{P}$ defined by, for $v\in\{0,1\}$, $\tilde{P}(U=1 \mid V=v)=0.9$, $\tilde{P}(U=2 \mid V=v)=0.09$. Then $\tilde{P}$ would still be safe for $\langle \mathbf{1}_{U=1}\rangle \mid \langle V\rangle$, but not for $\langle U\rangle \mid \langle V\rangle$: $\mathcal{P}^*$ contains a distribution whose marginal distribution $P(U=2)=0$, and (7) would not hold for that distribution.

Comparing the 'safety condition' (4) in Example 1 to (7) in Definition 1 we see that Definition 1 only imposes a requirement on expectations of U whereas (4) imposed a requirement also on RVs U' equal to functions $g(U)$ of U. For $\mathcal{U}$ with more than two elements as in Example 4, such a requirement is strictly stronger. We now proceed to define this stronger notion formally.

Definition 2 (*Stronger Safety Notions*)**.** Let $\mathcal{Z}$, $\mathcal{P}^*$, U, V and $\tilde{P}$ be as above. We say that $\tilde{P}$ is *safe* for $U|[\![V]\!]$ if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U'\rangle|[\![V]\!]$.

Similarly, $\tilde{P}$ is safe for $U \mid \langle V\rangle$ if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U'\rangle|\langle V\rangle$, and $\tilde{P}$ is safe for $U \mid [V]$ if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U'\rangle|[V]$.

We see that safety of $\tilde{P}$ for $U \mid [V]$ implies that $\mathrm{E}_{\tilde{P}}[g(U) \mid V=v]$ is the same for all values of v in the support of $\tilde{P}$, and all functions g of U. This can only be the case if $\tilde{P}(U \mid V)$ *ignores* V, i.e. $\tilde{P}(U \mid V=v)=\tilde{P}(U)$, for all supported v. We must then also have that, for all $v\in \mathrm{SUPP}_{\tilde{P}}(V)$, that $\tilde{P}(U)=P(U)$, which means that all distributions in $\mathcal{P}^*$ agree on the marginal distribution of U, and $\tilde{P}(U)$ is equal to this marginal distribution. Thus, $\tilde{P}$ is safe for $U \mid [V]$ iff it is *marginally valid*. A prime example of such a $\tilde{P}(U \mid V)$ that ignores V and is marginally correct is the $\tilde{P}(U \mid V)$ we encountered in Example 1.

To get everything in place, we need a final definition.

Definition 3 (*Safety Conditional on* W)**.** Let $\mathcal{Z}$, $\mathcal{P}^*$, U, V and $\tilde{P}$ be as above, and let W be another generalized RV.

1. We say that $\tilde{P}$ is *safe* for $\langle U\rangle|[\![V]\!]$, W if for all $w\in \mathrm{SUPP}_{\tilde{P}}(W)$, $\tilde{P} \mid W=w$ is safe for $\langle U\rangle \mid [\![V]\!]$ relative to $\mathcal{P}^* \mid W=w$. We say that $\tilde{P}$ is *safe* for $U|[\![V]\!]$, W if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U'\rangle|[\![V]\!]$, W.
2. The same definitions apply with $[\![V]\!]$ replaced by $\langle V\rangle$ and $[V]$.
3. We say that $\tilde{P}$ is *safe* for $\langle U\rangle|W$ if it is safe for $\langle U\rangle|[\![\mathbf{0}]\!]$, W; it is safe for $U|W$ if it is safe for $U|[\![\mathbf{0}]\!]$, W.

These definitions simply say that safety for '..|.., W' means that the space $\mathcal{Z}$ can be partitioned according to the value taken by W, and that for each element of the partition (indexed by w) one has 'local' safety given that one is in that element of the partition.

Proposition 1 gives reinterpretations of some of the notions above. The first one, (9) will mostly be useful for the proof of other results; the other three serve to make the original definitions more transparent:

Proposition 1 (*Basic Interpretations of Safety*)**.** *Consider the setting above. We have:*

1. $\tilde{P}$ *is safe for* $U|\langle V\rangle$ *iff for all* $P \in \mathcal{P}^*$, *there exists a distribution* P' *on* $\mathcal{Z}$ *with for all* $(u, v) \in \text{RANGE}((U, V))$, $P'(U = u, V = v) = \tilde{P}(U = u \mid V = v) \cdot P(V = v)$, *that satisfies*

$$P'(U) = P(U). \tag{9}$$

2. $\tilde{P}$ *is safe for* $\langle U\rangle|V$ *iff for all* $P \in \mathcal{P}^*$,

$$E_P[U \mid V] =_P E_{\tilde{P}}[U \mid V]. \tag{10}$$

3. $\tilde{P}$ *is safe for* $U \mid V$ *iff for all* $P \in \mathcal{P}^*$,

$$P(U \mid V) =_P \tilde{P}(U \mid V). \tag{11}$$

4. $\tilde{P}$ *is safe for* $U \mid [V], W$ *iff for all* $P \in \mathcal{P}^*$,

$$P(U \mid W) =_P \tilde{P}(U \mid V, W). \tag{12}$$

Together with the preceding definitions, this proposition establishes the arrows in Fig. 1 from $U \mid V$ to $\langle U\rangle \mid V$ (since (11) implies (10)), from $\langle U\rangle \mid V$ to $\langle U\rangle \mid \langle V\rangle$ and from $U \mid \langle V\rangle$ to $\langle U\rangle \mid \langle V\rangle$ (since (10) and (9) both imply (7)). The remaining arrows will be established by Theorems 1 and 2, under which we also discuss the characterizations on the right of the figure.

Note that (12) says that $\tilde{P}$ is safe for $U \mid [V], W$ if $\tilde{P}$ ignores V *given* W, i.e. according to $\tilde{P}$, U is conditionally independent of V given W. Thus, $\tilde{P}$ can be safe for $U \mid [V], W$ and still $\tilde{P}(U \mid V)$ may depend on V; the definition only requires that V is ignored once W is given.

(11) effectively expresses that $\tilde{P}(U \mid V)$ is valid (a frequentist might say 'true') for predicting U based on observing V, where as always we assume that $\mathcal{P}^*$ itself correctly describes our beliefs or potential truths (in particular, if $\mathcal{P}^* = \{P\}$ is a singleton, then any $\tilde{P}(U \mid V)$ which coincides a.s. with $P(U \mid V)$ is automatically valid). Thus, 'validity for $U \mid V$', to be interpreted as $\tilde{P}$ *is a valid distribution to use when predicting* U *given observations of* V is a natural name for safety for $U \mid V$. We also have a natural name for safety for $\langle U\rangle \mid V$: for 1-dimensional U, (10) simply expresses that all distributions in $\mathcal{P}^*$ agree on the conditional expectation of $U \mid V$, and that $\mathrm{E}_{\tilde{P}}[U \mid V]$ is a version of it. This implies (see e.g. Williams (1991)) that, with the function $g(v) := \mathrm{E}_{\tilde{P}}[U \mid V = v]$,

$$\mathrm{E}_{(U,V)\sim P}[(U - g(V))^2] = \min_f \mathrm{E}_{(U,V)\sim P}[(U - f(V))^2], \tag{13}$$

the minimum being taken over all functions from $\text{RANGE}(V)$ to $\mathbb{R}$. This means that $\tilde{P}$ encodes the *optimal* regression function for U given V and hence suggests the name *squared-error optimality*. Summarizing the names we encountered (see Fig. 1):

Definition 4 (*(Potential) Validity, Squared Error-Optimality, Unbiasedness, Marginal Validity*)**.** If $\tilde{P}$ is safe for $U|V$, i.e. (11) holds for all $P \in \mathcal{P}^*$, then we also call $\tilde{P}$ *valid* for $U \mid V$ (again, pronounce as "valid for *predicting* U *given* V"). If (11) holds for *some* $P \in \mathcal{P}^*$, we call $\tilde{P}$ *potentially valid* for $U \mid V$. If $\tilde{P}$ is safe for $\langle U\rangle|V$, we call $\tilde{P}$ *squared error-optimal* for $U \mid V$. If $\tilde{P}$ is safe for $\langle U\rangle|\langle V\rangle$, we call $\tilde{P}$ *unbiased* for $U \mid V$. If $\tilde{P}$ is safe for $\langle U\rangle \mid [V]$, we say that it is *marginally valid* for $U \mid V$.

It turns out that there also is a natural name for safety for $U \mid [V], W$ whenever $V \rightsquigarrow W$. The next example reiterates its importance, and the next section will provide the name: *calibration*.

Example 5. Suppose $\tilde{P}$ is safe for $U \mid [V_1], V_2$. From Proposition 1, (12) we see that this means that for all $P \in \mathcal{P}^*$, all $v_1, v_2 \in \text{SUPP}_P(V_1, V_2)$, that

$$\mathrm{E}_P[U \mid V_2 = v_2] = \mathrm{E}_{\tilde{P}}[U \mid V_1 = v_1, V_2 = v_2], \tag{14}$$

The special case with $V_2 \equiv \mathbf{0}$ has already been encountered in Example 1, (3). As discussed in that example, for $V_2 \equiv \mathbf{0}$, (14) expresses our basic interpretation of safety that *predictions based on* $\tilde{P}$ *will always be as good, in expectation, as the DM who uses* $\tilde{P}$ *expects them to be*. Clearly this continues to be the case if (14) holds for some nontrivial V_2.

2.3. Calibration safety

In this section, we show that *calibration*, as informally defined in Example 2, has a natural formulation in terms of our safety notions. We first define calibration formally, and then, in our first main result, Theorem 1, show how being calibrated for predicting U based on observing V is essentially equivalent to being safe for $U \mid [V], V'$ for some types of V' that need not be equal to V itself, including $V' \equiv \mathbf{0}$. Thus, we now effectively unify the ideas underlying Example 1 (dilation) and Example 2 (calibration).

Following Grünwald and Halpern (2011) we define calibration directly in terms of distributions rather than empirical data, in the following way:

Definition 5 (*Calibration*). Let $\mathcal{Z}, \mathcal{P}^*, U, V$ and $\tilde{P}$ be as above. We say that $\tilde{P}$ is *calibrated* (or *calibration–safe*) for $\langle U \rangle \mid V$ if for all $P \in \mathcal{P}^*$, all $\mu \in \{\mathrm{E}_{\tilde{P}}[U \mid V = v] : v \in \mathrm{SUPP}_P(V)\}$,

$$\mathrm{E}_P[U \mid \ \mathrm{E}_{\tilde{P}}[U \mid V] = \mu \ ] = \mu. \tag{15}$$

We say that $\tilde{P}$ is *calibrated* for $U \mid V$ if for all $P \in \mathcal{P}^*$, all $p \in \{\tilde{P}(U \mid V = v) : v \in \mathrm{SUPP}_P(V)\}$,

$$P(U \mid \ \tilde{P}(U \mid V) = p \ ) = p. \tag{16}$$

Hence, calibration (for U) means that given that a DM who uses $\tilde{P}$ *predicts* a specific distribution for U, the *actual* distribution is indeed equal to the predicted distribution. Note that here we once again treat $\tilde{P}(U \mid V)$ as a generalized RV.

In practice we would want to weaken Definition 5 to allow some slack, requiring the μ (viz. p) inside the conditioning to be only within some $\epsilon > 0$ of the μ (viz. p) outside, but the present idealized definition is sufficient for our purposes here. Note also that the definition refers to a simple form of calibration, which does not involve selection rules based on past data such as used by, e.g., Dawid (1982).

We now express calibration in terms of our safety notions. We will only do this for the 'full distribution'-version (16); a similar result can be established for the average-version.

Theorem 1 (*Calibration and Safety*). *Let U, V and $\tilde{P}$ be as above. The following three statements are equivalent:*

1. *$\tilde{P}$ is calibrated for $U|V$;*
2. *There exists a RV V' on $\mathcal{Z}$ with $V \rightsquigarrow_{\tilde{P}} V'$ such that $\tilde{P}$ is safe for $U \mid [V], V'$.*
3. *$\tilde{P}$ is safe for $U \mid V''$ where V'' is the generalized RV given by $V'' \equiv \tilde{P}(U \mid V)$.*

Note that, since safety for $U \mid V$ implies safety for $U \mid [V], V'$ for $V' = V$, (2.) $\Rightarrow$ (1.) shows that safety for $U \mid V$ implies calibration for $U \mid V$. By mere definition chasing (details omitted) one also finds that (2.) implies that $\tilde{P}$ is safe for $U \mid \langle V \rangle, V'$ and, again by definition chasing, that $\tilde{P}$ is safe for $U \mid \langle V \rangle$. Thus, this result establishes two more arrows of the hierarchy of Fig. 1. Its proof is based on the following simple result, interesting in its own right:

Proposition 2. *Let V and V' be generalized RVs such that $V \overset{f}{\rightsquigarrow}_{\tilde{P}} V'$ for some function f. The following statements are equivalent:*

1. *$\tilde{P}(U \mid V, V')$ ignores V, i.e. $\tilde{P}(U \mid V, V') =_{\tilde{P}} \tilde{P}(U \mid V')$.*
2. *For all $v' \in \mathrm{SUPP}_{\tilde{P}}(V')$, for all $v \in \mathrm{SUPP}_{\tilde{P}}(V)$ with $f(v) = v'$: $\tilde{P}(U \mid V = v) = \tilde{P}(U \mid V' = v')$.*
3. *$V' \rightsquigarrow_{\tilde{P}} \tilde{P}(U \mid V)$.*
4. *$V' \rightsquigarrow_{\tilde{P}} V''$ and $\tilde{P}(U \mid V', V'')$ ignores V', where $V'' = \tilde{P}(U \mid V)$.*

Moreover, if $\tilde{P}$ is safe for $U \mid V$ and $\tilde{P}(U \mid V, V')$ ignores V, then $\tilde{P}$ is safe for $U \mid V'$.

3. Continuous-valued U and V; confidence and pivotal safety

Our definitions of safety were given for countable $\mathcal{Z}$, making all random variables involved have countable range as well. Now we allow general $\mathcal{Z}$ and hence continuous-valued U and general uncountable V as well, but we consider a version of safety in which we do not have safety for $U \mid V$ itself, but for $U' \mid V$ for some U' with $(U, V) \rightsquigarrow U'$ such that the range of $\tilde{\mathcal{P}}_{[V]}(U') := \{\tilde{P}(U' \mid V = v) : v \in \mathrm{RANGE}(V)\}$ is still countable. To make this work we have to equip $\mathcal{Z}$ with an appropriate σ-algebra $\Sigma_{\mathcal{Z}}$ and have to add to the definition of a RV that it must be measurable,[5] and we have to modify the definition of support $\mathrm{SUPP}_P(U)$ to the standard measure-theoretic definition (which specializes to our definition (5) whenever there exists a countable $\mathcal{U}$ such that $P(U \in \mathcal{U}) = 1$). Yet nothing else changes and all previous definitions and propositions can still be used.[6]

Additional notations and assumptions In this section we frequently refer to (cumulative) distribution functions of 1-dimensional RVs, for which we introduce some notation: for distribution $P \in \mathcal{P}^*$ and RV $W : \mathcal{Z} \to \mathbb{R}$, let $F_{[W]}(\cdot)$ denote the distribution function of W, i.e. $F_{[W]}(w) = P(W \leq w)$. The notation is extended to conditional distribution functions: for given $\tilde{P}(U \mid V)$, we let $\tilde{F}_{[U|V]}(u|v) := \tilde{P}(U \leq u \mid V = v)$. The subscripts $[W]$ and $[U|V]$ indicate the RVs under consideration; we will omit them if they are clear from the context. Note that we can consider these distribution functions as RVs: for all $z \in \mathcal{Z}$, $F_{[W]}(W)(z) = P(W \leq W(z))$ and $\tilde{F}_{[U|V]}(U|V)(z) = \tilde{P}(U \leq U(z) \mid V = V(z))$.

Since this greatly simplifies matters, we will often assume that either $P \in \mathcal{P}^*$ or $\tilde{P}(U \mid V)$ satisfy the following:

Scalar density assumption A distribution $P(U)$ for RV U satisfies the *scalar density assumption* if (a) $\mathrm{RANGE}(U) \subseteq \mathbb{R}$ is equal to some (possibly unbounded) interval, and (b) P has a density f relative to Lebesgue measure with $f(u) > 0$ for all u in

[5] Formally we assume that $\mathcal{Z}$ is equipped with some σ-algebra $\Sigma_{\mathcal{Z}}$ that contains all singleton subsets of $\mathcal{Z}$. We associate the co-domain of any function $X : \mathcal{Z} \to \mathbb{R}^k$ with the standard Borel σ-algebra on $\mathbb{R}^k$, and we call such X an RV whenever the σ-algebra $\Sigma_{\mathcal{Z}}$ on $\mathcal{Z}$ is such that the function is measurable.

[6] If we were to consider safety of the form $U \mid V$ for uncountable $\tilde{\mathcal{P}}_{[V]}(U)$, then this set of probability distributions would have to be equipped with a topology, which is a bit more complicated and is left for future work.

the interior of RANGE(U). We say that $P(U|V)$ satisfies the scalar density assumption if for all $v \in$ RANGE(V), $P(U \mid V = v)$ satisfies it.

This is a strong assumption which will nevertheless be satisfied in many practical cases. For example, normal distributions, gamma distributions with fixed shape parameter, beta distributions etc. all satisfy it.

Overview of this section The goal of the following two subsections is to precisely reformulate the *fiducial* and *confidence* distributions that have been proposed in the statistical literature as pragmatic distributions in our sense, that can be safely used for some ('confidence-related') but not for other prediction tasks. Here we focus on the standard statistical scenario introduced in Example 3. The underlying idea of 'pivotal safety' (developed in Section 3.2) has applications in discrete, nonstatistical settings as well, on which we will report in another paper — we briefly return to them in Section 4.

3.1. Confidence safety

We start with a classic motivating example.

Example 6 (*Example 3, Specialized*)**.** As a special case of the statistical scenario outlined in Example 3, let $\mathcal{M}$ be the normal location family with varying mean θ and fixed variance, say $\sigma^2 = 1$, and let $V := \hat{\theta} = \hat{\theta}(X^n)$ where $\hat{\theta}(X^n)$ is the empirical average of the X_i, which is a sufficient statistic that is of course also equal to the ML estimator for data X^n. Then the sampling density of $\hat{\theta}$ is itself Gaussian, and given by

$$p(\hat{\theta} \mid \theta) \propto q_\theta(X^n) \propto e^{-\frac{1}{2}\cdot n\cdot(\hat{\theta}-\theta)^2}. \tag{17}$$

In this simple context, Fisher's controversial fiducial reasoning amounts to observing that (17) is symmetric in $\hat{\theta}$ and θ; thus, if we simply define a new function $\tilde{p}(\theta \mid \hat{\theta}) := p(\hat{\theta} \mid \theta)$, then this function must, for each fixed $\hat{\theta}$, be the density of a probability distribution (the integral over θ must by symmetry be 1); and this would then amount to something like a 'prior–free' posterior for θ based on data $\hat{\theta}$. In this special case, as well as with the corresponding inversion for scale families, it coincides with the Bayes posterior based on an improper Jeffreys' prior. Yet, Lindley (1958) showed that the general construction for 1-dimensional families, which we review in the next subsection, *cannot* correspond to a Bayesian posterior except for location and scale families: for different sample sizes, the 'fiducial' posterior for data of size n corresponds to the Bayes posterior for a prior which depends on n.

Fisher (1930) noted that $\tilde{p}$ as constructed above lead to valid inference about confidence intervals. Later (Fisher, 1935) he made claims that $\tilde{p}$ could be used for general prior-free inference about θ given data/statistic $\hat{\theta}$. This is not correct though, and more recently, $\tilde{p}$ is more often regarded as an instance of a *confidence distribution* (Schweder and Hjort, 2002), a term going back to Cox (1958) — these are by and large the same objects as fiducial distributions, though with a stipulation that they only be used for certain inferences related to confidence. In the remainder of this subsection, we develop a variation of safety that can capture such confidence statements. In the next subsection, we review the general method for designing confidence distributions for 1-dimensional statistical families and we shall see that, under an additional condition, they are indeed confidence-safe in our sense. In the remainder of this section, we focus on 1-dimensional families and interpret the RVs U and V as in our statistical application of Examples 3 and 6. Thus, $U \equiv \theta$ would be a 1-dimensional scalar parameter of some model $\{P_\theta : \theta \in \Theta\}$, $V \equiv S(X^n)$ would be a statistic of the observed data. In Example 6, $V \equiv \hat{\theta}(X^n)$ is the ML estimator.

We are thus interested in constructing, for each $v \in$ RANGE(V), an interval of $\mathbb{R}$ that has (say) 95% probability under $\tilde{P}(U \mid V = v)$. To this end, we define for each $v \in$ RANGE(V), an interval $\mathcal{C}_v = [\underline{u}_v, \bar{u}_v]$ where $\underline{u}_v$ is such that $\tilde{F}_{[U|V]}(\underline{u}_v \mid v) = 0.025$ and $\bar{u}_v$ is such that $\tilde{F}_{[U|V]}(\bar{u}_v \mid v) = 0.975$. This set obviously has 95% probability according to $\tilde{P}(U \mid V = v)$. In our interpretation where $U = \theta$ is the parameter of a statistical model, we may interpret $\tilde{P}$ as DM's assessment, given data $V = S(X^n)$, of the uncertainty about U, i.e. $\tilde{P}$ is a 'posterior' and, analogous to Bayesian terminology, we may call $\mathcal{C}_v$ a 95% *credible set* given V. The question is now under what conditions we have *coverage*, i.e. that $\mathcal{C}_V$ is also a 95% frequentist *confidence interval*, so that our credible set can be given frequentist meaning. By definition of confidence interval, this will be the case iff for all $P \in \mathcal{P}^*$, $P(U \in \mathcal{C}_V) = 0.95$, i.e. iff for all $P \in \mathcal{P}^*$, $v \in$ RANGE(V),

$$\mathrm{E}_P[\mathbf{1}_{U\in\mathcal{C}_V}] = \mathrm{E}_{\tilde{P}}[\mathbf{1}_{U\in\mathcal{C}_V} \mid V = v], \tag{18}$$

where we used that, by construction, $\mathrm{E}_{\tilde{P}}[\mathbf{1}_{U\in\mathcal{C}_V} \mid V = v] = 0.95$ for all $v \in$ RANGE(V). As we shall see (18) holds for our normal example, so the posterior constructed in (17) produces valid confidence intervals. (18) is of the form of a 'safety' statement and it suggests that confidence interval validity of credible sets can be phrased in terms of safety in general. Indeed this is possible as long as $\tilde{P}(U|V)$ satisfies the scalar density assumption: for fixed $0 \le a < b \le 1$, we can define the set $\mathcal{C}_v^{[a,b]} = [\underline{u}_v^a, \bar{u}_v^b]$ where $\tilde{F}_{[U|V]}(\underline{u}_v^a \mid v) = a$ and $\tilde{F}_{[U|V]}(\bar{u}_v^b \mid v) = b$, so that for each $v \in$ RANGE(V), $\mathcal{C}_v^{[a,b]}$ is a $b - a$ credible set. Reasoning like above, we then get that $\mathcal{C}_V^{[a,b]}$ is also a $b - a$ confidence interval iff for all $P \in \mathcal{P}^*$, all $v \in$ RANGE(V)

$$\mathrm{E}_{(U,V)\sim P}[\mathbf{1}_{U\in\mathcal{C}_V^{[a,b]}}] = \mathrm{E}_{\tilde{P}}[\mathbf{1}_{U\in\mathcal{C}_V^{[a,b]}} \mid V = v], \tag{19}$$

which, from the characterization of safety for $U \mid [V]$, Proposition 1, (12) and (14) suggests the following definition:

Definition 6 (*Confidence Safety*)**.** Let U, V and $\tilde{P}$ be such that $\tilde{P}(U|V = v)$ satisfies the scalar density assumption for all $v \in \text{RANGE}(V)$. We say that $\tilde{P}$ is (strongly) *confidence-safe* for $U \mid V$ if for all $0 \leq a < b \leq 1$, it is safe for $\mathbf{1}_{U \in \mathcal{C}_V^{[a,b]}} \mid [V]$.

The requirement that $\tilde{P}$ satisfies the scalar density assumption is imposed because otherwise $\mathcal{C}_V^{[a,b]}$ may not be defined for some $a, b \in [0, 1]$. We could also consider distributions that have coverage in a slightly weaker sense, and define weak confidence-safety for $U \mid V$ as safety for $\mathbf{1}_{U \in \mathcal{C}_V^{[a,b]}} \mid \langle V \rangle$; we have not (yet) found any natural examples though that exhibit weak confidence-safety but not strong confidence safety.

Example 7 (*Safe and Unsafe Decisions based on Confidence Distributions*)**.** In the next subsection we show that $\tilde{P}(\theta \mid V = \hat{\theta}(X^n))$ as defined in Example 6 (normal distributions) is confidence-safe. For example, we may specify a $\tilde{P}(\theta \mid \hat{\theta}) - 95\%$ credible set $\mathcal{C}_V^{[a,b]}$ with $a = 0.025$ and $b = 0.975$ as the area under the normal curve centered at $V = \hat{\theta}$ and truncated so that the area under the left and right remaining tails is 0.025 each. Now suppose that $X^n \sim P_\theta$ for arbitrary θ. By confidence-safety we know that the probability that we will observe $\hat{\theta}$ such that $\theta \notin \mathcal{C}_V^{[a,b]}$ is exactly 0.05, just as it would be if $\tilde{P}$ where the true conditional distribution — an instance of a *safe* inference based on $\tilde{P}$. For an example of an inference that is *unsafe*, suppose DM really is offered a gamble for \$1 that pays out \$2 whenever $\theta > 0$ (we could take any other fixed value as well), and pays out 0 otherwise. She thus has two actions at her disposal, $a = 1$ (accept the gamble) and $a = 0$ (abstain), with loss given by $L(\theta, 0) = 0$ for all θ and $L(\theta, 1) = 1$ if $\theta < 0$ and $L(\theta, 1) = -1$ otherwise. She might thus be tempted to follow the decision rule $\delta(\hat{\theta})$ that accepts the gamble whenever she observes $\hat{\theta}$ such that $\tilde{P}(\theta > 0 \mid \hat{\theta}) > .5$ and abstains otherwise; for that rule minimizes, among all decision rules, her expected loss $\mathrm{E}_{\theta \sim \tilde{P}|\hat{\theta}}[L(\theta, \delta(\hat{\theta}))]$, which is nonpositive and even strictly negative whenever $\tilde{P}(\theta > 0 \mid \hat{\theta}) > .5$.

This decision rule is problematic though, because it is based on an inference that is not safe in any of our senses: safety would mean that $\tilde{P}$ is safe for $L(\theta, \delta(\hat{\theta})) \mid \mathrm{s}$, where s can be substituted by $[\hat{\theta}]$, $\langle \hat{\theta} \rangle$, or $\hat{\theta}$. The first does not apply since $\hat{\theta}$ is not ignored in the probability assessment; the third does not hold because it would imply the second, which also does not hold. To see this, note that if data comes from P_θ with $\theta < 0$ then we have

$$\mathrm{E}_{\hat{\theta} \sim P_\theta}[L(\theta, \delta(\hat{\theta}))] > 0 > \mathrm{E}_{\hat{\theta} \sim P_\theta}[\mathrm{E}_{\theta \sim \tilde{P}|\hat{\theta}}[L(\theta, \delta(\hat{\theta}))]], \tag{20}$$

so that her actual expected loss is positive whereas she thinks it to be negative. This latter conclusion – the rightmost inequality in (20) – follows because, as we established earlier, her expected loss $\mathrm{E}_{\theta \sim \tilde{P}|\hat{\theta}}[L(\theta, \delta(\hat{\theta}))]$ is nonpositive and strictly negative on an event that has θ-probability > 0. (20) violates (7) in Definition 1 so that $\tilde{P}$ is not safe for $L(\theta, \delta(\hat{\theta})) \mid \langle \hat{\theta} \rangle$. Note that, if $\tilde{P}$ were safe for $\theta \mid \hat{\theta}$ (as a subjective Bayesian would believe if $\tilde{P}$ were her posterior) then it would also be safe for $L(\theta, \delta(\hat{\theta}))$ (because $L(\theta, \delta(\hat{\theta}))$ can be written as a function of $(\theta, \hat{\theta})$), and then use of δ would be safe after all.

For an intuitive interpretation, consider a long sequence of experiments. For each j, in the jth experiment, a sample of size $n = 10$ is drawn from a normal with some mean θ_j. Each time DM investigates whether $\theta_j > 0$. Assume that, in reality, all or most of the θ_j are < 0, but DM does not know this. Then every once in a while $\hat{\theta}$ will be large enough for our unsafe DM to gamble on it, but every time this happens she loses; all other times she neither loses nor wins, so her net gain is negative in the long run whereas she expects it to be positive. On the other hand, it might also be the case that in reality, nearly all of the θ_j are much larger than 0. Then everytime she accepts the gamble, she actually gains 1, although she herself expects to gain (somewhat) less: in some cases, $P(\theta > 0 \mid \hat{\theta})$ will be only slightly larger than 0.5 and, while she expects to gain only slightly more than 0.5 on average in those cases, she will in fact always win. Here we treat both cases (gaining more than one expects and gaining less than one expects) as *unsafe* — although one could envision 'one-sided' notions of safety in which only gaining less is defined to be 'unsafe'.

Note also that nothing would change in principle if the loss for action $a = 0$ would be set to a small but constant negative value — so that DM would be guaranteed to gain money by always abstaining. If she played the decision rule that minimized her posterior expected loss, she would still sometimes take the gamble though, and her actual long-run gain might again be different from what she expects it to be.

Thus, $\tilde{P}$ is not safe for $\theta \mid \hat{\theta}$ in general. However, it is still safe for $U'|V$ for some other functions of $U \equiv \theta$ besides $U' = \mathcal{C}_V^{[a,b]}$. For example, it leads to unbiased estimation of the mean: $\tilde{P}$ is safe for $\langle \theta \rangle \mid \langle \hat{\theta} \rangle$, as is easily established. This is however a special property of the confidence distribution for the normal location family when θ denotes the mean — the unbiasedness is not preserved under 1-to-1 reparameterizations of the parameter (see Example 9) and also – not surprisingly – fails to hold for the more general 1-dimensional confidence distributions, which we review below.

3.2. Pivotal safety and confidence

Trivially, if $\tilde{P}$ is safe for $U|V$ (hence valid) and the scalar density assumption holds, then it is also confidence-safe for $U|V$. We now determine a way to construct confidence-safe $\tilde{P}$ if not enough knowledge is available to infer a $\tilde{P}$ that is valid. To this end, we invoke the concept of a *pivot*, usually defined as a function of the data and the parameter that has the same distribution for every $P \in \mathcal{P}^*$ and that is monotonic in the parameter for every possible value of the data (Barndorff-Nielsen and Cox, 1994). We adopt the following variation that also covers a quite different situation with discrete outcomes, relating to puzzles such as Monty Hall — we will report on these in a different paper.

Definition 7 (*Pivot*)**.** Let U and V be as before and suppose either (continuous case) that $U : \mathcal{Z} \to \mathbb{R}$ and $V : \mathcal{Z} \to \mathbb{R}$ are real-valued RVs, and that for all $v \in \text{RANGE}(V)$, $\text{RANGE}(U|V = v)$ is a (possibly unbounded) interval (possibly dependent on v), or (discrete case) that $\mathcal{Z}$ is countable. We call RV U' a (continuous viz. discrete) *pivot* for $U \mid V$ if

1. $(U, V) \rightsquigarrow U'$ so that the function f with $U' = f(U, V)$ exists.
2. For each fixed $v \in \text{RANGE}(V)$, the function $f_v : \text{RANGE}(U|V = v) \to \text{RANGE}(U')$, defined as $f_v(u) := f(u, v)$ is 1-to-1 (an injection); in the continuous case we further require f_v to be continuous and uniformly monotonic, i.e. either $f_v(u)$ is increasing in u for all $v \in \text{RANGE}(V)$, or $f_v(u)$ is decreasing in u for all $v \in \text{RANGE}(V)$.
3. All $P \in \mathcal{P}^*$ agree on U', i.e. for all $P_1, P_2 \in \mathcal{P}^*$, $P_1(U') = P_2(U')$, where in the continuous case we further require that P_1 (hence also P_2) satisfies the scalar density assumption.

We say that a pivot U' is *simple* if for all $v \in \text{RANGE}(V)$, the function f_v is a bijection.

The scalar density assumption (item 3) does not belong to the standard definition of pivot, but it is often assumed implicitly, e.g. by Schweder and Hjort (2002). The importance of 'simple' pivots (a nonstandard notion) will become clear below.

In the remainder of this section we focus on the statistical case of the previous subsection, which is a special case of Definition 7 — thus invariably $U \equiv \theta$, the 1-dimensional parameter of a model $\{P_\theta \mid \theta \in \Theta\}$, and V is some statistic of data X^n. In Section 4 we very briefly return to the discrete case.

If a continuous pivot as above exists, then all $P \in \mathcal{P}^*$ have the same distribution function $F_{[U']}(u') := P(U' \leq u')$. We may thus define a pragmatic distribution by setting, for all $v \in \text{RANGE}(V)$, all $u \in \text{RANGE}(U \mid V = v)$,

$$\tilde{F}_{[U|V]}(u \mid v) := \begin{cases} F_{[U']}(f_v(u)) & \text{if } f_v(u) \text{ increasing in } u \\ 1 - F_{[U']}(f_v(u)) & \text{if } f_v(u) \text{ decreasing.} \end{cases} \tag{21}$$

The definition of pivot ensures that for each $v \in \text{RANGE}(V)$, $\tilde{F}_{[U|V]}(u \mid v)$ is a continuous increasing function of u that is in between 0 and 1 on all $u \in \text{RANGE}(U \mid V = v)$, and hence $\tilde{F}_{[U|V]}(u \mid v)$ is the CDF of some distribution $\tilde{P}(U|V)$. It can be seen from the standard definition of a confidence distribution (Schweder and Hjort, 2002) that this $\tilde{P}(U|V)$ is a confidence distribution, and that every confidence distribution can be obtained in this way.[7] Hence, (21) essentially defines confidence distribution. Theorem 2 shows that when based on *simple* pivots, confidence distributions are also confidence-safe.

Example 8 (*Simplification for Exponential Families*)**.** Consider the statistical setting with $U \equiv \theta$, $V \equiv \hat{\theta}(X^n)$, and (a) for all $\theta \in \Theta$, $P_\theta(V)$ itself satisfies the scalar density assumption, and (b) for each fixed $v \in \text{RANGE}(V)$, we have that $F_{\theta,[V]}(v) := P_\theta(\hat{\theta}(X^n) \leq v)$ is monotonically decreasing in θ. This will hold for 1-dimensional exponential families with a continuously supported sufficient statistic (such as the normal, exponential, beta- and many other models), taken in their mean-value parameterization Θ. Then (by (b)) $U' = F_{\theta,[V]}(V)$ is itself a decreasing pivot, with (by (a)) the additional property that the function f_θ from $\text{RANGE}(V)$ to $\text{RANGE}(U')$ given by $f_\theta(v) := f(\theta, v)$ is strictly increasing in v. Then (21) simplifies, because (using this strict increasingness in the second equality):

$$F_{\theta,[V]}(v) = P_\theta(V \leq v) = P_\theta(F_{\theta,[V]}(V) \leq f_\theta(v)) = F_{\theta,[F_{\theta,[V]}]}(f_\theta(v)) = F_{[U']}(f_v(\theta)),$$

and noticing that the right-hand side appears in (21), we can plug in the left-hand side there as well and we see that we can directly set

$$\tilde{F}(\theta \mid \hat{\theta}) = 1 - F_\theta(\hat{\theta}). \tag{22}$$

Thus for such models the recipe (21) simplifies (see also Veronese and Melilli (2015)).

We now define 'pivotal safety' which, as demonstrated below, in the statistical case essentially coincides with confidence safety — the reason for the added generality is that it also has meaning and repercussions in the discrete case. The extension to 'multipivots' is just a stratification that means that, given any $w \in \text{RANGE}(W)$, $\tilde{P} \mid W = w$ is pivotally safe for $U \mid V$ relative to $\mathcal{P}^* \mid W = w$; it is not really needed in this text, but is convenient for completing the hierarchy in Fig. 1.

Definition 8 (*Pivotal Safety*)**.** Let U and V be as before and let $\tilde{P}$ be an arbitrary distribution on $\mathcal{Z}$ (not necessarily given by (21)). If V has full support under $\tilde{P}$, i.e. $\text{SUPP}_{\tilde{P}}(V) = \text{RANGE}(V)$ and U' is a (continuous or discrete) pivot such that $\tilde{P}$ is safe for $U' \mid [V]$, i.e. for all $v \in \text{RANGE}(V)$,

$$\tilde{P}(U' \mid V = v) = \tilde{P}(U'),$$

then we say that $\tilde{P}$ is pivotally safe for $U \mid V$, with pivot U'.

Now let W be a generalized RV such that $V \rightsquigarrow W$. Suppose that for all $w \in \text{RANGE}(W)$, U' is a pivot relative to the set of distributions $(\mathcal{P}^* \mid W = w)$ and $\tilde{P}$ is safe for $U' \mid [V], W$. Then we say that $\tilde{P}$ is pivotally safe for $U \mid V$ with *multipivot* $U|W$.

[7] Mirroring the discussion underneath Definition 1 from Schweder and Hjort (2002): if $\tilde{F}'(U|V)$ is the CDF of a confidence distribution as defined by them, then $U' := \tilde{F}'(U|V)$ is a pivot and then the construction above applied to U' gives $\tilde{F}(U|V) := \tilde{F}'(U|V)$. Conversely, if U' is an arbitrary continuous pivot, then by the requirement that $P(U')$ has a density with interval support, $F(U')$ is itself uniformly distributed on its support $[0, 1]$ and there is a 1-to-1 continuous mapping between U' and $F(U')$. Thus, whenever U' is a continuous pivot, $F(U')$ is itself a pivot as well, and $\tilde{F}(U|V)$ as defined here satisfies the definition of confidence distribution.

Example 9 (*Normal Distributions and General Confidence Distributions*)**.** In Example 6, $U = \theta$, the mean of a normal with variance 1, and we set $V = \hat{\theta}(X^n)$ to be the average of a sample of size n. Then it is easily seen that $U' = \theta - \hat{\theta} = U - V$ is a pivot according to our definition, having a $N(0, 1)$ distribution under all $P \in \mathcal{P}^*$. If we adopt $\tilde{P}(U \mid V)$, under which $U' \sim N(0, 1)$ independently of V, then $\tilde{P}$ is safe for $U' \mid [V]$ (see Proposition 1, (12)) so we have pivotal safety. And indeed one can verify that this $\tilde{P}$ coincides with the recipe given by (21).

As an aside, we note that if we reparameterize a model in a smooth 1-to-1 fashion, then the recipe given by (21) leads to the same distribution which will still be confidence-safe (see also the rightmost column in Fig. 1 in the pivotal-safety row). However, other safety properties of $\tilde{P}$ may change under reparameterization. In our current normal location example, if θ represents the mean, $\tilde{P}$ is also unbiased (safe for $\langle\theta\rangle \mid \langle\hat{\theta}\rangle$), but if we reparameterize to $\theta' := \theta^3$, then $\tilde{P}(\theta' \mid V = v)$ is not unbiased any more, i.e. not safe for $\langle\theta^3\rangle \mid \langle\hat{\theta}^3\rangle$: we have $\mathrm{E}_{P_\theta}[\theta'] = \theta^3$, yet an easy calculation shows that, at sample size $n = 1$, $E_{\hat{\theta}|P_\theta}[E_{\theta\sim\tilde{P}|\hat{\theta}}[\theta']] = 6\theta^3 + 6\theta$.

In the simplest form of calibration, safety for $U \mid [V]$, it held that $\tilde{P}(U \mid V)$ was the marginal of U, so that U and V are independent under $\tilde{P}$. In the simplest pivotal safety case, the situation is comparable, but now the auxiliary variable U' instead of the original variable U is independent of V under $\tilde{P}$. Note though that we do not necessarily have that U' and V are independent for all or even for *any* $P \in \mathcal{P}^*$. In the statistical application below, we even have that $P(U' \mid V)$ is a degenerate distribution (putting all its mass on a particular real number depending on V) under each $P \in \mathcal{P}^*$, as can be checked from Example 6.

The relation between pivotal, confidence-safety and the $\tilde{P}$ defined as in (21) is given by the following central result.

Theorem 2 (*Characterizing Pivotal and Confidence Safety*)**.** *The following statements are all equivalent:*

1. *$\tilde{P}$ is pivotally safe for $U|V$ with some simple continuous pivot U'*
2. *$\tilde{P}(U|V)$ is of form (21), where U' is a simple continuous pivot*
3. *$\tilde{P}$ is confidence-safe for $U|V$ and for each $v \in \textsc{range}(V)$, $\tilde{P}(U|V = v)$ satisfies the scalar density assumption*
4. *$\tilde{P}$ is safe for $\tilde{F}(U|V)|V$ and for each $v \in \textsc{range}(V)$, $\tilde{P}(U|V = v)$ satisfies the scalar density assumption*
5. *$\tilde{P}$ is pivotally safe for $U|V$ with pivot $\tilde{F}(U|V)$, which is continuous and simple.*

The theorem shows that, whenever pivots are simple (as is often the case), confidence distributions $\tilde{P}$ as defined by (21) are also confidence-safe. If a pivot is nonsimple however, confidence distributions can still be defined via (21) but they may either not be confidence-safe or not satisfy the scalar density assumption — without which confidence-safety is undefined under our current definition. We now give an example of the latter:

Example 10 (*Non-simple Pivot*)**.** Let $\mathcal{M}$ be the 1-dimensional normal family, but the parameter of interest is now $U := |\theta|$ rather than θ. As shown by Schweder and Hjort (2016, Example 6.1), the confidence distribution $\tilde{P}(U \mid \hat{\theta})$ defined by (21) then gives a point mass $\tilde{P}(U = 0 \mid \hat{\theta} = v) = p_v > 0$ to $U = |\theta| = 0$ whose size p_v depends on v. This happens because $\tilde{F}(U \mid v)$ ranges, for such v, not from 0 to 1 but from p_v to 1, where for $v \downarrow 0$, $p_v \uparrow 0.5$. Then pivotal safety cannot be achieved for $\tilde{P}(U|V)$, since there must be v_1, v_2 with $\tilde{F}(U|V = v_1) \neq \tilde{F}(U|V = v_2)$ whence the definition is not satisfied. Now such confidence distributions based on nonsimple pivots are still useful, and indeed we can prove a different form of confidence safety for such cases — let us give a rough sketch: in general, when $\tilde{P}(U \mid V = v)$ has point masses, then $\tilde{F}(U \mid V = v)$ has discontinuities and hence $C_v^{[a,b]}$ may not be well-defined for some a and b, even though $0 \leq a < b \leq 1$. One solution is to allow the credible-quantile interval $[a_v, b_v]$ to depend on the observed v in such a way that for every observed v, $C_v^{[a_v,b_v]}$ is well-defined — we should just make sure that a_v and b_v are continuity points of $\tilde{F}(U \mid V = v)$. We must now be very careful however, because for such confidence-intervals with data-dependent quantiles, confidence distributions may not necessarily be confidence-safe any more! However, if we define a_v and b_v in a special way then we still have a different safety property. Namely, fix some $0 \leq a < b \leq 1$ and set for each v, a_v to be the largest number $\leq a$ and b_v to be the smallest number $\geq b$ such that $C_v^{[a_v,b_v]}$ is well-defined. Then the confidence-distribution $\tilde{P}$ as defined by Schweder and Hjort (2016) is *safe for* $\mathbf{1}_{U \in C_V^{[a_V,b_V]}} \mid [\![V]\!]$. Thus, viz. (6), we have for all $P \in \mathcal{P}^*$:

$$b - a \leq \inf_{v \in \mathrm{SUPP}_{\tilde{P}}(V)} \tilde{P}(U \in C_V^{[a_V,b_V]} | V = v] \leq P(U \in C_V^{[a_V,b_V]}) \leq \sup_{v \in \mathrm{SUPP}_{\tilde{P}}(V)} \tilde{P}(U \in C_V^{[a_V,b_V]} | V = v].$$

Thus, we still get a safe estimate of confidence if, no matter what actual v we observe, we take as the probability of the observed confidence interval the *smallest* probability of the confidence interval given any v. In the present example, if we are interested in, say, a 90% confidence interval we would like to set $a = 0.05, b = 0.95$. But if we observe a small v' (so that $\tilde{P}(U = 0 \mid V = v') > 0.05$), we may be forced to set $a_{v'} = 0, b_{v'} = 0.95$ and hence $\tilde{P}(U \in C_{v'}^{[a_{v'},b_{v'}]} \mid V = v') = 0.95$. We may now be tempted to claim that 'conditionally on observing small v', we get a better (95%) confidence interval', but then we are moving on unsafe terrain: the notion of safety $\mid [\![V]\!]$ only guarantees a 90% confidence, even if the specific v we observe suggests something else.

Comparing to Definition 8, we see that this type of safety for $\mid [\![V]\!]$ may be viewed as a conservative form confidence-safety; it needs further study, however, and we do not know whether an analogue of Theorem 2 holds for this notion — see Section 4.

The hierarchy To see how pivotal and confidence safety fit into the hierarchy of Fig. 1, note that Theorem 2 establishes the double arrow between pivotal safety and confidence safety under the scalar density assumption (SDA) — the requirement that $(U', V) \rightsquigarrow U$ in the figure amounts to f_v being a bijection, as we require. The theorem also establishes the relation between calibration and pivotal safety, under the assumption that $\tilde{P}(V)$ has full support and the SDA holds for U. Then the simplest form of calibration, safety for $U \mid [V]$, clearly implies pivotal safety for $U \mid V$ — just take $U' = U$, which is immediately checked to be a pivot. This result trivially extends to the general case of safety for $U \mid [V]$, V' with $V' \neq \mathbf{0}$, this implying pivotal safety with multipivot $U \mid V'$ — we omit the details.

It remains to establish the rightmost column of Fig. 1; we will only do this in an informal manner. Schweder and Hjort (2002) (and, implicitly, Hampel (2006)) already note that if $\tilde{P}$ is a confidence distribution for RV U given data V, then it remains a confidence distribution for monotonic functions U' of U, but not for general functions of U. In our framework this translates to, under the scalar density assumption of Section 3.1, that pivotal safety of $U \mid V$ implies pivotal safety for $U' \mid V$ if U' is a 1-to-1 continuous function of U, which readily follows from Definition 7 and Theorem 2 (Definition 7 implies an analogous statement for the discrete case as well). Similarly, it is a straightforward consequence from the definitions that calibration for $U \mid V$ implies calibration for $U' \mid V$, for every U' with $U \rightsquigarrow U'$, not necessarily 1-to-1; yet for U' with $(U, V) \rightsquigarrow U'$, calibration may not be preserved: take e.g. the setting of Example 1 (dilation) with $U' = |V - U|$. Then $\tilde{P}(U' = 1 \mid V = 0) = 0.9$, $\tilde{P}(U' = 1 \mid V = 1) = 0.1$, yet $\mathcal{P}^*$ contains a distribution with $P(U = V) = 1$ and for this P, $P(U' = 1 \mid V) \equiv 0$. If $\tilde{P}$ is valid for $U \mid V$ however, validity is preserved even for every U' with $(U, V) \rightsquigarrow U'$.

4. Earlier and future work; open problems and conlusion

Earlier work and the most important future work The idea that fiducial or confidence distributions can be used for valid assessment of some, not all, RVs or events that can be defined on a domain has been stressed by several authors, e.g. Schweder and Hjort (2002), Xie and Singh (2013) and Hampel (2006). The novelty here is that we formalize the idea and place it in broader context and hierarchy. The idea of replacing sets of distributions by a single representative also underlies the MDL Principle (de Rooij and Grünwald, 2011), yet again, without broader context or hierarchy. It is also the core of the pignistic transformation advocated by Smets (1989) as part of his *transferable belief model*, which, apart from the transformation idea, seems to be almost totally different from safe probability however — it would be interesting to sort out the relations though. I already noted in the introduction that my own earlier work contains various definitions of partial notions of safety, but unifying concepts, let alone a hierarchy, were lacking.

There is one crucial issue though that we neglected in this paper, and that was brought up earlier, to some extent, by Grünwald (2000) and Grünwald and Halpern (2004): the fact that mere safety is not enough — we want our pragmatic $\tilde{P}$ to also have optimality properties (see e.g. Example 2 for the trivial weather forecaster who is calibrated (safe) without being optimal). Some interesting results are given by Schweder and Hjort (2016) for confidence-safety: they provide several theorems that, in our language, establish when a particular confidence-safe distribution is optimal within the set of all confidence-safe distributions for the given task. We also need to know, however, when 'safe' (in some respect) distributions are also optimal (in some sense) among the set of *all* distributions one might want to use for making predictions, not just the safe ones. As indicated by Grünwald (2000) and more implicitly by van Ommen et al. (2016), there is a link between safety and minimax optimality which sometimes can be exploited, but much remains to be done here — this is our main goal for future work.

Additional work II: pivotal safety and probability puzzles Closely related to the developments in this paper, and explicitly involving loss functions and optimality notions, is the application of pivotal safety to probability puzzles such as the Monty Hall problem. This work has been completed, but we will report on it elsewhere. Let us briefly indicate the issues here using Monty Hall as an example. Key to understanding this problem is the realization that, when the contestant has directly chosen the door with the car behind it, then the quiz master has a *choice* — he can open either of the other two doors. In most analyses of the Monty Hall problem that are usually viewed as 'correct', one implicitly assumes that the quiz master flips a *fair* coin to decide which of these two doors to open. There have been heated discussions (e.g. on wikipedia talk pages) about whether this assumption is justified. Now, one can show that the $\tilde{P}$ which assumes a fair coin flip by Monty is an instance of a *pivotally safe* pragmatic distribution. Such distributions, we prove, have the property that for many loss functions (including all symmetric loss functions, such as the 0/1-loss as in Monty Hall), they lead one to making optimal decisions. Thus, while assuming a fair coin flip may be wrong, it is in many – but not all! – cases still *safe* to base one's decisions upon it.

Additional work III: event-based conditioning We can think of our pragmatic $\tilde{P}(U|V)$ as probability updating rules, mapping observations $V = v$ to distributions on U. We can thus extend our approach to update distributions given *events* rather than RVs, leading, for example, to the 'sanity check' for the Monty Hall problem that we announced in the introduction. As a special case, we obtain that when the set of events that one might observe is not a partition of the sample space, then conditioning on such events is *unsafe*, in the sense that it violates an analogue of our safety definition. The fact that conditioning is problematic if one conditions on something not equal to a partition has in fact been known for a long time (see e.g. Shafer (1985) for the first landmark reference, and see Grünwald and Halpern (2003) for broader context). The point here is merely that it nicely fits into the safety framework; this work has also been completed and will be reported on elsewhere.

Please cite this article in press as: Grünwald P., Safe probability. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.09.014.

Earlier and future work IV: misspecification Suppose we do inference based on statistical model $\mathcal{M}$ and assume that the data are sampled i.i.d. from some P^* that is not necessarily in $\mathcal{M}$. Bayesian and other likelihood-methods, if they converge at all, will then typically converge to the *reverse-information-projection* of P^* onto $\mathcal{M}$, which coincides with the distribution $\tilde{P}$ in $\mathcal{M}$ that minimizes the KL divergence $D(P^* \parallel P)$ among all $P \in \mathcal{M}$, whenever such a minimizer exists. Since in general $\tilde{P} \neq P^*$, predictions based on $\tilde{P}$ may be wrong or suboptimal, and we may study for what prediction tasks $\tilde{P}$ is still useful. It turns out that, for some models $\mathcal{M}$, one can guarantee various types of *safety* for some types of predictions; for example, in regression with a homoskedastic Gaussian noise model, the elements $P_f(Y|X)$ of $\mathcal{M}$ express that $Y = f(X) + \epsilon$ where $\epsilon \sim N(0, 1)$. If P^* is such that $\mathcal{M}$ contains P_{f^*} for the true regression function f^* but the true noise ϵ, while 0-mean, is not Gaussian, then $P^* \neq \tilde{P}$ but one can still show that $\tilde{P} = P_{f^*}$ is safe for $\langle Y \rangle | X$ in our notation. Thus, here one takes $\mathcal{P}^*$ to be the set of *all* distributions on (X, Y) with regression function f^*, i.e. including distributions in which ϵ is not Gaussian. Grünwald and van Ommen (2014) establish several more of such safety properties depending on the model $\mathcal{M}$ and its degree of misspecification. They call random variables for which a safety guarantee holds *model-associated prediction tasks*, building on the paper *the Safe Bayesian* (Grünwald, 2012). While this work has already been done, it was not phrased explicitly in the safety language of this paper, and an interesting question for future work is whether one can establish a general theory of precisely what prediction tasks are safe in precisely what sense for what models under what further conditions on P^*.

One can also analyze the situation one level higher and consider 'posterior' distributions $\tilde{P}(U \mid V)$ as in Example 3. Thus while above, $\tilde{P}$ was a distribution on data, it is now a conditional distribution on parameters U (or distributions) given data V. The misspecified case extends Example 3 by not assuming that $\mathcal{P}^* \equiv \mathcal{M}$; instead one takes $\mathcal{P}^*$ to be strictly larger. One can then establish whether one's inference methods are safe at this higher level, in various respects. One may, for example, view hypothesis testing with optional stopping (see below) as an instance of our framework where one small but important aspect of the model $\mathcal{M}$ is misspecified: namely, $\mathcal{M}$ assumes a fixed number of data points, whereas $\mathcal{P}^*$ contains distributions in which the sample size N is itself a random variable. A robust p-value as mentioned below then satisfies a form of safety which implies that Type-I error guarantees remain valid under optional stopping.

Future work V: safe testing There is one application of safe probability that is particularly promising, and we hope to finish a paper on it soon. This is the use of safety concepts in *testing*. As is well-known, the notion of p-value is wrought with practical problems — it does not deal well with optional stopping, it depends on aspects of the sampling plan that may be unknown or unknowable, and so on (Wagenmakers, 2007). Bayesian methods resolve such issues to some extent, but not fully (van der Pas and Grünwald, 2018). It turns out that all these issues can be rephrased in terms of 'safety' — and one can define, for example, 'robust' p-values (van der Pas and Grünwald, 2018) that lead to Type-I error guarantees that remain valid under optional stopping. Standard p-values can then be reinterpreted as pragmatic probabilities that are safe (in the $|[\![\cdot]\!]$ sense) for a (small) class of inferences; robust p-values are safe for a larger class of inferences.

Future work VI: objective bayes Safe Probability may also be fruitfully applied to objective Bayesian inference. For example, consider inference of a Bernoulli parameter based on an 'objective' Jeffreys' prior $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$. Use of such a prior may certainly be defensible because of its geometric and information-theoretic properties (de Rooij and Grünwald, 2011), but what if we have a very small sample of just 1 or even 0 outcomes? Then Jeffreys' prior would tell us, for example, that a bias θ between 0 and 0.01 is 10 times as likely than a bias between 0.495 and 0.505. Most objective Bayesians would probably not be prepared to gamble on that proposition.[8] This is fine, but then what propositions would an objective Bayesian be prepared to gamble on, and what not? Bayesian inference has no tools to deal with this question – and – in a manner similar to characterization of safety for fiducial distributions — safe probability may offer them.

Future work VII: epistemic probability More generally, both objective Bayesian and fiducial methods have been proposed as candidates for *epistemic probability* (Keynes, 1921; Carnap, 1950; Hampel, 2006) but it is unclear how exactly such a notion of probability should be connected to decision theory — while a Bayesian or frequentist probability of 0.01 on outcome A implies that a (not too risk-averse) DM would be willing to pay one dollar for a lottery ticket that pays off 200 dollar if A turns out to be the case, for epistemic probability this is not so clear. Safe probability suggests that it might be fruitful to view epistemic probabilities as assuming a willingness to bet on a *strict subset* of all events $\mathcal{A}$ that can be defined on the given space.

Future work VIII: extending confidence safety — handling all types of confidence distributions Our work on confidence distributions in Section 3.1 only dealt with the simplest and cleanest case of confidence distributions, namely those where both the distributions in $\mathcal{P}^*$ and the confidence distribution itself satisfy the scalar density assumption. Confidence distributions can be usefully defined in various other situations — for example, they may be improper; or the confidence sets may not be single intervals (Schweder and Hjort, 2016, Chapter 4). Also, for discrete data, approximate confidence distributions may be defined. We briefly encountered one such non-simple situation (point mass on some θ) in Example 10. The example suggests that our notion of confidence safety can be extended to this and (we suspect) other nonstandard

[8] One might object that an actual value of θ may not even exist, and certainly will never be observed, so one cannot gamble on it. But I could propose this gamble instead: I will toss the biased coin 10 000 times, and only reveal to you the final relative frequency of heads. How much would you bet on it being ≤ 0.01?

cases, but much work needs to be done here. Of particular interest is the question, raised by the example, whether confidence safety is retained if the credible set $C_v^{[a_v,b_v]}$ output by the inference is allowed to depend on v in some mild manner.

Open problems Other future work involves open problems, as mentioned in the caption of Fig. 1. Of particular interest is whether we can extend confidence safety to multidimensional U. Earlier work (Dawid and Stone, 1982; Seidenfeld, 1992) suggests that then in general, there will be multiple, different choices for $\tilde{P}$, none of which is inherently 'best'. A major additional goal for future work is to identify subjective considerations that may lead one to prefer one choice over another, cf. the idea of 'luckiness' (de Rooij and Grünwald, 2011). Another intriguing question is whether safety can be re-constructed as an *extension* of measure theory — which has also been designed to restrict the notion of (probability) measures so that they cannot just be applied to any set one likes. This would help to clarify the notion of epistemic probability. Yet another avenue is to extend the definition of pragmatic distributions using upper- and lower expectations, replacing $\tilde{P}$ by a set of distributions $\tilde{\mathcal{P}}$ (this is briefly detailed in Appendix A.1). Then both $\tilde{\mathcal{P}}$ and $\mathcal{P}^*$ would fall into the 'imprecise probability' paradigm; we could still get nontrivial predictions as long as $\tilde{\mathcal{P}}$ is more 'specific' than $\mathcal{P}^*$. Such an extension would hopefully allow us to represent the random-set approach to fiducial inference from Dempster (1968) and its modern extensions, such as the inferential models of Martin and Liu (2013), as an extension of pivotal safety. Here confidence-safe probabilities would be replaced by confidence-safe probability intervals; perhaps one could even arrive at a general description of what applications of Dempster–Shafer theory (Shafer, 1976; Dempster, 1968) are safe at all, and if so, to what degree they are safe.

Acknowledgments

This manuscript has benefited a lot from various discussions over the last fifteen years with Philip Dawid, Joe Halpern and Teddy Seidenfeld. Special thanks go to Teddy as well as to Gert de Cooman and Nils Hjort for providing encouragement that was essential to get this work done. This research was supported by the Netherlands Organization for Scientific Research (NWO) VICI Project Nr. 639.073.04.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jspi.2017.09.014.

References

Augustin, T., Coolen, F.P.A., de Cooman, G., Troffaes, M.C.M., 2014. Introduction to Imprecise Probabilities. John Wiley & Sons.
Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Chapman and Hall.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Box, G.E.P., 1979. Robustness in the strategy of scientific model building. In: Launer, R.L., Wilkinson, G.N. (Eds.), Robustness in Statistics. Academic Press, New York.
Carnap, R., 1950. Logical Foundations of Probability. University of Chicago Press.
Cox, D.R., 1958. Some problems connected with statistical inference. Ann. Math. Statist. 29 (2), 357–372.
Dawid, A.P., 1982. The well-calibrated Bayesian. J. Amer. Statist. Assoc. 77, 605–611 Discussion: pages 611–613.
Dawid, A.P., Stone, M., 1982. The functional-model basis of fiducial inference. Ann. Statist. 1054–1067.
de Rooij, S., Grünwald, P.D., 2011. Luckiness and regret in minimum description length inference. In: Bandyopadhyay, Prasanta S., Forster, M. (Eds.), Handbook of the Philosophy of Science, Vol. 7. Elsevier.
Dempster, A.P., 1968. A generalization of Bayesian inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 205–247.
Efron, B., 1996. R.A. Fisher in the 21st century: invited paper presented at the 1996 R.A. Fisher lecture. Statist. Sci. 13 (2), 95–122.
Fisher, R.A., 1930. Inverse probability. Proc. Cambridge Philos. Soc. 26, 528–535.
Fisher, R.A., 1935. The fiducial argument in statistical inference. Ann. Eugenics 6, 391–398.
Foster, D.P., Vohra, R.V., 1998. Asymptotic calibration. Biometrika 85 (2), 379–390.
Fraser, D.A.S., 1968. The Structure of Inference, Vol. 23. Wiley, New York.
Fraser, D.A.S., 1979. Inference and Linear Models. McGraw-Hill, New York.
Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. J. Math. Econom. 18 (2), 141–153.
Gill, R.D., 2011. The Monty Hall problem is not a probability puzzle — it's a challenge in mathematical modelling. Statist. Neerlandica 65 (1), 58–71.
Grünwald, P.D., 1999. Viewing all models as probabilistic. In: Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99), pp. 171–182.
Grünwald, P.D., 2000. Maximum entropy and the glasses you are looking through. In: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence. (UAI 2000), Morgan Kaufmann, San Francisco, pp. 238–246.
Grünwald, P.D., 2012. The safe Bayesian. In: Proceedings of the 23rd International Conference on Algorithmic Learning Theory (ALT '12), Lyon, France.
Grünwald, P.D., 2013. Safe probability: restricted conditioning and extended marginalization. In: Proceedings Twelfth European Conference on Symbolic and Quantitative Approaches To Reasoning with Uncertainty. (ECSQARU 2013), In: Lecture Notes in Computer Science, vol. 7958, Springer, pp. 242–252.
Grünwald, P.D., Halpern, J.Y., 2003. Updating probabilities. J. Artificial Intelligence Res. 19, 243–278.
Grünwald, P.D., Halpern, J.Y., 2004. When ignorance is bliss. In: Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2004), Banff, Canada.
Grünwald, P.D., Halpern, J.Y., 2011. Making decisions using sets of probabilities: Updating, time consistency, and calibration. J. Artif. Intell. Res. (JAIR) 42, 393–426.
Grünwald, P.D., van Ommen, T., 2014. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. Arxiv Preprint 1412.3730.
Hampel, F., 2001. An outline of a unifying statistical theory. In: ISIPTA, pp. 205–212.
Hampel, F., 2006. The proper fiducial argument. In: Ahlswede, R. (Ed.), Information Transfer and Combinatorics. In: LNCS, Springer Verlag, pp. 512–526.

Hannig, J., 2009. On generalized fiducial inference. Statist. Sinica 491–544.
Keynes, J.M., 1921. Treatise on Probability. Macmillan, London.
Kolmogorov, A.N., 1933. Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer-Verlag.
Lindley, D.V., 1958. Fiducial distributions and Bayes' theorem. J. R. Stat. Soc. Ser. B Stat. Methodol. 102–107.
Martin, R., Liu, C., 2013. Inferential models: A framework for prior-free posterior probabilistic inference. J. Amer. Statist. Assoc. 108 (501), 301–313.
Pearl, J., 2009. Causality, second ed. Cambridge University Press.
Ramsey, F.P., 1931. Truth and probability (1926). The foundations of mathematics and other logical essays, pp. 156–198.
Schweder, T., Hjort, N.L., 2002. Confidence and likelihood. Scand. J. Statist. 29 (2), 309–332.
Schweder, T., Hjort, N., 2016. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge University Press.
Seidenfeld, T., 1992. R.A Fisher's fiducial argument and Bayes' theorem. Statist. Sci. 358–368.
Seidenfeld, T., Wasserman, L., 1993. Dilation for convex sets of probabilities. Ann. Statist. 21, 1139–1154.
Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton University Press.
Shafer, G., 1985. Conditional probability. Internat. Statist. Rev. 53 (3), 261–277.
Smets, P., 1989. Constructing the pignistic probability function in a context of uncertainty. In: UAI, Vol. 89, pp. 29–40.
Sweeting, T.J., 2001. Coverage probability bias, objective bayes and the likelihood principle. Biometrika 88 (3), 657–675.
Taraldsen, G., Lindqvist, B.H., 2013. Fiducial theory and optimal inference. Ann. Statist. 41 (1), 323–341.
van der Pas, S., Grünwald, P., 2018. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. Statist. Sinica 28 (1).
van Ommen, T., Koolen, W.M., Feenstra, T.E., Grünwald, P.D., 2016. Updating probability beyond conditioning on a partition. Internat. J. Approx. Reason. 74, 30–57.
Veronese, P., Melilli, E., 2015. Fiducial and confidence distributions for real exponential families. Scand. J. Statist. 42 (2), 471–484.
Vos Savant, M., 1990. Ask Marilyn. Parade Magazine, page 15, There were also followup articles in Parade Magazine on Dec. 2, 1990 (p. 25) and Feb. 17, 1991 (p. 12).
Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic Learning in a Random World. Springer, New York.
Wagenmakers, E.J., 2007. A practical solution to the pervasive problems of p values. Psychon. Bull. & Rev. 14 (5), 779–804.
Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. In: Monographs on Statistics and Applied Probability, vol. 42, Chapman and Hall, London.
Williams, D., 1991. Probability with Martingales. Cambridge Mathematical Textbooks.
Xie, M., Singh, K., 2013. Confidence distribution, the frequentist distribution estimator of a parameter: a review. Internat. Statist. Rev. 81 (1), 3–39.

Please cite this article in press as: Grünwald P., Safe probability. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.09.014.