On the likelihood of "encapsulating all uncertainty".

Kristy A Martire[1,2]*, Gary Edmond[3], Danielle J Navarro[1] & Ben R Newell[1]

1. School of Psychology, University of New South Wales
2. ORCID ID: orcid.org/0000-0002-5324-0732
3. Faculty of Law, University of New South Wales orcid.org/0000-0003-2609-7499

*Corresponding author: Kristy A Martire; School of Psychology, The University of New South Wales, Sydney, NSW, Australia, 2052; k.martire@unsw.edu.au

Conflicts of interest: none.

Citation:

KA Martire, G Edmond, DJ Navarro and BR Newell (2017). On the likelihood of 'encapsulating all uncertainty' *Science and Justice, 57*, 76-79.

https://doi.org/10.1016/j.scijus.2016.10.004

Abstract

The assignment of personal probabilities to form a forensic practitioner's likelihood ratio is a mental operation subject to all the frailties of human memory, perception and judgment. While we agree that beliefs expressed as coherent probabilities are neither 'right' nor 'wrong' we argue that debate over this fact obscures both the requirement for and consideration of the 'helpfulness' of practitioner's opinions. We also question the extent to which a likelihood ratio based on personal probabilities can realistically be expected to 'encapsulate all uncertainty'. Courts cannot rigorously assess a forensic practitioner's bare assertions of belief regarding evidential strength. At a minimum, information regarding the uncertainty both within and between the opinions of practitioners is required.

1. Introduction

From our reading of the position papers in this special issue, there are two different conceptualizations of the nature of a forensic practitioner's likelihood ratio. Consistent with the frequentist approach to statistical inference, some authors take the position that the practitioner's likelihood ratio should be "an empirically calculated estimate of a true but unknown value" [1]. This perspective emphasises the fact that orthodox point estimates of an unknown quantity are subject to sampling variability, and some measure of this variability should therefore accompany the likelihood ratio. In the scientific literature the Neyman confidence interval [2] is commonly used for this purpose. In contrast, other authors take a Bayesian perspective and argue that probabilities reflect the epistemic uncertainty of an intelligent reasoner. As such, these authors argue that the "likelihood ratio" should be a Bayes factor[1] representing "personal viewpoints" [3] that express a "rational degree of belief". [4] To the extent that such beliefs emerge from the application of a properly formulated Bayesian statistical model (one that can be made available to the trier of fact), we do not take issue with the claim that probabilities can express epistemic uncertainty, and that these probabilities may provide assistance to the court. However, to the extent that reported likelihood ratios are the product of experience and introspection of actual humans rather than of idealised reasoners (e.g., Jaynes' hypothetical "Bayesian robot"; [5]), some care is required. As scholars of human behaviour and decision making we will restrict our response to issues relating to the uncertainty relevant to the latter type of likelihoods. In Section 2 of this response we consider the issue of the 'warrant' for opinions of personal belief [1], exploring the implications of *misleading* (rather than 'right' or 'wrong') subjective probabilities. In Section 3 we examine the validity of the claim that the (any) likelihood ratio

[1] Like other commentators we note that Bayes factors and likelihood ratios are very different statistical concepts, but will use the terms interchangeably for the purposes of this article.

"encapsulates all uncertainty" [4]. Finally, in Section 4 we present our conclusions.

2. What is the warrant for opinions expressing a rational degree of belief?

Like Morrison and Enzinger [1], and Risinger [6] before them, we are of the opinion that likelihood ratios sincerely stating subjective beliefs [7-10] raise problems for courts and may not satisfy admissibility standards in jurisdictions requiring reliability[2] [11]. The presentation of impression, beliefs or "guesses" [6] is "the opposite of what is desired at trial; the primary goal is objectively true results" [13, p. 10. See also, 14]. Moreover, if forensic practitioners endorsing the presentation of personal probabilities truly believe that there is "no 'right' or 'true' probability" [3], and that there is consequently no way to assign a probability that is "'more likely to be close to the *right* probability'" [3], it is unclear how they believe their opinion evidence can assist the court to obtain objectively true results. Resolving this fundamental disconnect between what courts expect and what some practitioners are offering seems central to regulating the admission and use of expert opinion evidence into the future.

One factor potentially contributing to the apparent chasm between legal expectations and statements of subjective belief may be the use of terms such as 'true' and 'right' (and their antonyms). Proponents of subjective probabilities are quick to emphasise that the probabilities that form the basis of a likelihood ratio (LR) are neither "known or unknown" [3], they are not "something real that exists in the external world" [4] and they do "not pretend to get the 'true' probabilities" [9]. Thus there can be no consideration of the accuracy[3] of the practitioner's subjective belief, because there is no ground truth against which to assess its veracity. The LR is what the practitioner believes, and what they believe can't be 'wrong'. We agree with elements of this argument, but contend that it ultimately distracts from more important issues. Courts (including judges and jurors) do not necessarily expect that a forensic practitioner will only ever provide objectively correct, true or right opinions (as the frequent admission of expert witnesses with opposing opinions demonstrates), they do however expect to be provided with information that will help them to evaluate the facts in issue [15]. This includes whether and to what extent forensic opinions of this type will tend to assist or impair the accurate resolution of a dispute in issue.

[2] Our reference here is to *evidentiary* reliability – that is, trustworthiness as per [11] Daubert v. Merrell Dow Pharmaceuticals, Inc, in: US, Supreme Court, 1993, pp. 579. See also [12] IMM v. The Queen, in: ALJR, 2016, pp. 529: "within the scheme of the Evidence Act, evidence that is trustworthy is evidence that is 'reliable'" [12] IMM v. The Queen, in: ALJR, 2016, pp. 529.

[3] Here the term 'accuracy' refers to the closeness of the assigned value to the true (unknown) value of the likelihood ratio.

We argue that focusing on misunderstandings regarding the 'rightness' and 'wrongness' of personal probability statements obscures the more fundamental issue: the extent to which the expressed opinion is *helpful* or *misleading* to the court. Indeed, we would happily accept the statement that personal probabilities are neither accurate nor inaccurate if we could then open up a dialogue about ways to estimate and communicate the potential for personal probabilities to mislead lawyers, judges and juries. To date, engagement by practitioners with the misleading potential of personal probability opinions has been limited [8, 16] even though there is acknowledgement that the criminal justice system uses the opinions of practitioners to inform decision making [10], and that misleading evidence should be minimised [13].

This limited engagement with the potential for misleading opinions may stem from an assumption evident in this special issue and in the literature more generally. Notwithstanding the fact that a personal probability is neither objectively right nor wrong, practitioners who endorse personal probabilities as the basis for LRs seem to believe that given enough experience and information, subjective probabilities will converge on better and more accurate descriptions of the underlying state of affairs (i.e., the true but unknown and unknowable value). For example: "… as more data is obtained, the use of subjective prior information becomes less relevant and different scientists will end up in close agreement" [4]. This being the case, estimates of the misleading potential of the practitioner's opinion are not necessary because given enough time and enough information, their beliefs will become helpful. Indeed, unless the practitioner's beliefs converge on truth, their opinions cannot assist the court to reach accurate outcomes.

While we would also anticipate that practitioner's opinions will converge on truth given relevant data, experience and training, we believe this can only occur if certain conditions hold. That is, we agree with the more specific statement that: "*individual beliefs will converge on the truth if updated over time with the objectively true results of repeated experiments that can be fashioned into objectively true conditional probabilities"* [13]. This leads us to question whether it is reasonable to believe that forensic practitioners are able to update their beliefs over time with 'objectively true results' in order to fashion them into 'objectively true conditional probabilities'[4] which 'encapsulate all uncertainty' [4].

3. Can a likelihood ratio "encapsulate all uncertainty"

Both Biedermann et al [3] and Berger and Slooten [4] argue that "there is no need to assign a measure of uncertainty to the measure of uncertainty" [3] that a likelihood ratio

[4] i.e., where the assigned conditional probabilities closely correspond to the true (but unknown) underlying value.

(or Bayes factor) describes. We partially agree. To the extent that the practitioner relies on a Bayesian approach to probability, it makes little sense to place a frequentist confidence interval around the LR, as this fundamentally misunderstands the epistemic claim implied by a Bayes factor. However, we disagree with any suggestion that this is the only respect in which "uncertainty" is relevant to the trier of fact.

As the papers by Ommen, Saunders & Neumann [17] and Taylor, Hicks & Champod [18] highlight, the probative value of a Bayesian analysis may be minimal if the reported LR is extremely sensitive to the choice of prior distribution. Courts may be loathe to admit testimony from an expert who applies an inappropriate model, or whose LR can be radically altered by only a modest change to the data available, regardless of the degree of certainty implied by their reported LR. Indeed, in applied Bayesian data analysis [19] it is common to undertake posterior predictive checks to verify that the model is consistent with the data it purports to explain, and to check the sensitivity to the prior to verify that one's conclusions are driven more by data than by prior biases. These practical considerations invariably require a scientist to disclose a good deal more about their thinking and expose more of the internal workings of their model than they might have done were the LR deemed to be the sole criterion upon which an analysis is to be judged. Furthermore, it may also require the scientist to disclose more information about their decision-making processes than they can accurately or explicitly know (we will discuss in more detail ahead).

To illustrate, consider the following example, loosely based on the simple beta binomial model in Example 2 of Berger and Slooten [B&S; 4]. A crime has been committed in Science City, and one of the pieces of physical evidence at the scene reveals that both the perpetrator and the suspect spell the word "*colour*" with an "u" (as per UK usage), rather than the customary "*color*" spelling used in Science City (as per US usage). How strong is this as evidence for the prosecution? Two Experts (A and B) are given access to the same database of writing samples, and use it in the following way. Both experts assume that there is some unknown probability p that a randomly chosen member of "the population" will use the UK spelling of the word "colour", and use a uniform prior (Beta distribution parameters a=b=1) to express their *a priori* uncertainty about the prevalence of the UK spelling of "colour". The two experts use the same prior, and they use the same beta-binomial model that gives rise to Equation 7 in B&S [4]. However, they apply different personal statistical models because they disagree about how the parameter p should be defined: Expert A decides that the physical evidence "colour" indicates that the perpetrator is originally from the United Kingdom and after consulting the 'population' database they discover m=98 of the people who provided writing samples are from the United Kingdom, and n=9899 people are not. At trial, Expert A applies

Equation 7 from the B&S paper to conclude that the LR is

LR = (1+1+98+1+9899) / (1 + 1 + 98) = 10000 / 100 = 100

Accordingly the testimony offered is an LR of 100:1 favoring the prosecution. In contrast Expert B construes the problem more narrowly, and decides that the only relevant data correspond to the spelling of the word color/colour specifically, because people from the United Kingdom living in Science City may choose to adopt the customary "color" spelling and people originating from outside the UK might also use the UK spelling (as in Australia). After consulting the same database, Expert B discovers that – due to different samples containing different text - there are only m=3 people known to use the "colour" spelling, and n=494 people known to use the "color" spelling. Again applying the beta-binomial model Expert B concludes – by a remarkable coincidence – that the LR is

LR = (1+1+3+1+494) / (1 + 1 + 3) = 500 / 5 = 100

At trial, Experts A and B offer exactly the same LR. Moreover, they have used the same statistical tool (the beta-binomial model), applied the same prior (uniform), and have used the same database. However, because they have interpreted the parameter of interest differently, their testimony about the "colour" observation pertains to subtly different quantities (i.e., the proportion of people from the UK vs the proportion of people who use the UK spelling of "colour") and draws on a different subset of the observations in the database. Given the differences in how the experts have interpreted the statistical problem, their approaches may not be equally *sensitive* to the arrival of new evidence.

For instance, suppose a new data set were made available to both experts, including 50 people that use the "colour" spelling, all of whom are originally from the UK and 50 people who use the "color" spelling, none of whom were from the UK. Both experts would deem this new data relevant, but would revise their beliefs in a very different way. Expert A would now set m=148 and n=9999, and revise the LR to be 67:1 favouring the prosecution but Expert B would revise the numbers to m=53 and n=544, yielding the not entirely convincing LR of 11:1. The important point here is that, in the absence of explicit information from the expert accurately describing the personal statistical model they applied, the fact finder cannot anticipate how easily or how much an expert might be inclined to adjust their testimony in light of additional evidence. In everyday parlance (though not statistical nomenclature) this kind of sensitivity *is* a form of practitioner uncertainty, and we argue – much as Taylor et al [18] do - that these factors are relevant to the court even though it is not captured by the LR.

Importantly, this difference in sensitivity is not caused by one expert or the other having access to a larger database (both experts were given the same resources), and cannot easily be resolved by the trier of fact asking simple questions about the *quantity* of data upon which the LR is based. Expert A has used more data than Expert B, but it is not clear that their testimony has greater probative value. Ultimately, the trier of fact must make a determination about which expert has interpreted the problem in the most appropriate fashion, even though they might not ever be provided the information about the personal statistical model that has been applied to derive that information. Therefore to the extent that the probative value of evidence depends on subjective (likely opaque) factors, the psychological facts about how humans form beliefs are relevant to the court.

With this in mind we now expand our focus, and ask what happens in the situation where an actual human is involved, rather than an idealised reasoner such as Jaynes' infamous Bayesian robot? While an LR may indeed be "a construct of the human mind" as Berger and Slooten [4] argue, not all beliefs constructed by the human mind can plausibly be characterised as personal probabilities in the sense required by Bayesian reasoning. As eminent cognitive scientist Daniel Kahneman cautions us, we should be wary of believing whatever comes to mind [20]. Other researchers highlight the dangers associated with what has been termed "metacognitive myopia" [21]. Metacognitive myopia is a short-sightedness regarding the origins and generating processes of (internal and external) data used as the basis of opinions and beliefs. The probability assignments made by forensic practitioners are those of humans subject to all of the same biases and heuristics as non-practitioners [22]. These biases include phenomena such as availability whereby salient or evocative information is more readily brought to mind, and attributed greater weight than other more representative but less available information. Primacy and recency effects mean that we are more likely to recall instances that occur either early or late in a sequence than those occurring in the middle for no reason other than their relative positions. These and other cognitive idiosyncrasies have ultimately led scientists to differentiate the human mind from veridical recording devices [23]. The human mind does not automatically store experiences to be replayed on demand. Storing, encoding and retrieving information, including the information relevant to the assignment of personal probabilities, is an imperfect and reconstructive process.

Given that few forensic practitioners are likely to be familiar with the diverse and counterintuitive range of cognitive phenomena identified in the biases and heuristics literature, it is difficult to imagine that the descriptions of the personal statistical models utilised by scientists to assign personal probabilities will routinely take into account all relevant uncertainties. Moreover, metacognitive myopia is a phenomenon that compounds the difficulty in identifying and quantifying uncertainty even where some general insight

into cognitive limitations is present. These difficulties then extend to the elicitation, explanation and evaluation of expert opinions by lawyers, judges and juries.

Beyond these memory factors, cognitive scientists have identified structural characteristics of practitioners' learning and working environments as potential contributors to uncertainty. In brief, the extent to which a personal probability opinion is likely to be helpful or misleading to a court is dependent on an awareness of the representativeness or bias in the sampled information and potential errors in any conclusions. This is only possible if there is a) a sufficiently predictable environment in which to learn, and b) adequate feedback to facilitate the identification of regularities in the environment [24]. Without these conditions "there is literally no necessary connection between subjective belief states and correct outcomes" [13] or beliefs and it remains to be seen whether these conditions hold for forensic practitioners who provide statements of personal probability.

Putting the difficulties associated with identifying and quantifying all of the relevant cognitive and environmental factors which might contribute to the uncertainty of an assigned probability to one side, we also believe that there are other uncertainties, which reside outside the LR that are of interest to courts. In this special issue [3, 4] and elsewhere [16, 25, 26], scholars in favour of personal probabilities have explained that given different data or the experiences of different examiners with different uncertainties, different likelihood ratios will be assigned for the same observations. The extent to which the LRs assigned by different practitioners agree or cluster can be considered as a measure of precision according to the definition in the introduction to this special issue [27] (although we appreciate that the conceptualisation of the points in the cluster as individual practitioner's opinions rather than the results of algorithms or statistical computations may not be familiar in this context). In cognitive science, the examination of the clustering of judgements made by individuals regarding the same observations, evidence or tests is called 'inter-rater reliability'.

Thus, while it may be true that an LR based on personal probabilities is an expression of uncertainty which does not require its own measure of uncertainty *at the level of the individual practitioner*, we believe there is additional uncertainty to be considered. Given current knowledge we are agnostic about how this uncertainty should most appropriately be described and communicated and note proposals for two-step evaluations [17] and qualifying text or statements [3] in this special issue. Irrespective, we believe that there is something relevant and important for courts to derive from information about whether, and the extent to which, the LRs of appropriately qualified practitioners, given the same observations (evidence), vary. Assuming practitioners are generally more likely to provide

helpful rather than misleading opinions (which may or may not be reasonable given the discussion above), close correspondence between practitioners' opinions can provide information about the range within which the probative value of the evidence might (truly) lie. Conversely, low correspondence between the opinions of practitioners may suggest that the value of the evidence is far from settled and perhaps should not be considered reliable, and therefore may not be able to assist with fact-finding. It is important to note, however, that while high agreement between practitioners may be *necessary* for courts to consider practitioners' opinions reliable, it is not *sufficient* to establish reliability *per se*. Practitioners' opinions must tend to be objectively helpful, rather than misleading, if the court is to benefit from their admission (i.e., the performance of subjective human assigner(s) also needs to be empirically evaluated).

4. Conclusions

While we do not endorse the 'warrant' for opinions based only on subjective probabilities (in their current form) we feel it is important to engage with proponents on their terms in order to facilitate discussion and to foster mutual understanding. Beyond issues of warrant we have doubts about the extent to which personal probabilities can be assumed to provide courts with helpful rather than potentially misleading information, and appropriately encapsulate relevant uncertainties. Courts are not tasked with accepting expert opinions as bare assertions of belief, and they cannot rationally challenge or evaluate opinions that are essentially statements of faith. Rather they are required to determine whether the opinion is one which is likely to be sufficiently reliable to rationally influence their assessment of facts in issue [15, 28, 29]. At present forensic practitioners provide little information that could assist courts with this assessment although there seems to be agreement in this special issue that such information is necessary [3]. Clearly there are logical and mathematical challenges which need to be examined and resolved, an aim this special issue commendably addresses. Yet, there are also conceptual issues to be considered. There is uncertainty in practitioner opinions that is unlikely to be incorporated into LRs due to (understandable) ignorance of cognitive influences and myopic tendencies on the part of human decision makers. There is also uncertainty (inconsistency) between the opinions of examiners. We believe these are important issues for forensic practitioners as they determine how to enhance the presentation of their opinions as likelihood ratios.

Acknowledgements

This work was supported by an Australian Research Council (ARC) DECRA Fellowship to KAM (DE140100183) an ARC Discovery Project (DP160101186) to BRN and an ARC Linkage Project to KAM and GE (LP160100008).

References

[1] G.S. Morrison, E. Enzinger, What should a forensic practitioner's likelihood ratio be?, Science & Justice, (2016).

[2] J. Neyman, Outline of a theory of statistical estimation based on the classical theory of probability, Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236 (1937) 333-380.

[3] A. Biedermann, S. Bozza, F. Taroni, C. Aitken, Reframing the debate: A question of probability, not of likelihood ratio, Science & Justice, (2016).

[4] C.E.H. Berger, K. Slooten, The LR does not exist, Science & Justice, (2016).

[5] E.T. Jaynes, Probability theory: The logic of science, Cambridge university press, 2003.

[6] D.M. Risinger, Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'), Law, Probability and Risk, (2012) mgs011.

[7] I.W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Science & Justice, 38 (1998) 198-202.

[8] I.W. Evett, G. Jackson, J. Lambert, More on the hierarchy of propositions: exploring the distinction between explanations and propositions, Science & Justice, 40 (2000) 3-10.

[9] F. Taroni, C.G.G. Aitken, P. Garbolino, De Finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: a commentary for forensic scientists, Science & Justice, 41 (2001) 145-150.

[10] A. Biedermann, P. Garbolino, F. Taroni, The subjectivist interpretation of probability and the problem of individualisation in forensic science, Science & Justice, 53 (2013) 192-200.

[11] Daubert v. Merrell Dow Pharmaceuticals, Inc, in: US, Supreme Court, 1993, pp. 579.

[12] IMM v. The Queen, in: ALJR, 2016, pp. 529.

[13] R.J. Allen, The nature of juridical proof: Probability as a tool in plausible reasoning, International Journal of Evidence and Proof, (in press).

[14] J. Becker, United States v. Downing, in: 753 F.2d 1224, 1985.

[15] G. Edmond, Forensic Science Evidence and the Conditions for Rational (Jury) Evaluation, Melb. UL Rev., 39 (2015) 77.

[16] F. Taroni, S. Bozza, A. Biedermann, C.G.G. Aitken, Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio, Law, Probability and Risk, 15 (2015) 1-16.

[17] D.M. Ommen, C.P. Saunders, C. Neumann, An argument against presenting interval quantifications as a surrogate for the value of evidence, Science & Justice, (2016).

[18] D. Taylor, T. Hicks, C. Champod, Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios, Science & Justice,

(2016).
[19] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian data analysis, Chapman & Hall/CRC Boca Raton, FL, USA, 2014.
[20] D. Kahneman, Thinking, fast and slow, Macmillan, 2011.
[21] K. Fiedler, 1 Meta-Cognitive Myopia and the Dilemmas of Inductive-Statistical Inference, Psychology of Learning and Motivation-Advances in Research and Theory, 57 (2012) 1.
[22] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, science, 185 (1974) 1124-1131.
[23] F.C. Bartlett, Remembering: An experimental and social study, Cambridge: Cambridge University, (1932).
[24] D. Kahneman, G. Klein, Conditions for intuitive expertise: a failure to disagree, American Psychologist, 64 (2009) 515.
[25] G. Jackson, P.J. Jones, Case assessment and interpretation, Wiley Encyclopedia of Forensic Science, (2009).
[26] A. Nordgaard, Comment on 'Dismissal of the illusion of uncertainty on the assessment of a likelihood ratio' by Taroni F., Bozza S., Biederman A. and Aitken C, Law, Probability and Risk, 15 (2015) 17-22.
[27] G.S. Morrison, Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate, Science & Justice, (2016).
[28] Davie v. Edinburgh Magistrates, in: SC, 1953, pp. 34.
[29] Makita (Australia) Pty Ltd v. Sprowles, in: NSWLR, 2001, pp. 705.