Law, Probability and Risk, 2024, **00**, 1–28
https://doi.org/10.1093/lpr/mgae007
Research Article

OXFORD

Decisionalizing the problem of reliance on expert and machine evidence

Alex Biedermann [1,*] **and Timothy Lau** [2,†]

[1]University of Lausanne, Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, 1015 Lausanne–Dorigny, Switzerland
[2]Research Division, Federal Judicial Center, Thurgood Marshall Federal Judiciary Building, Washington, DC, 20002-8003, United States

*Corresponding author. E-mail: alex.biedermann@unil.ch

†The views expressed in this article are of the author alone and do not represent the views of the Federal Judicial Center.

Abstract

This article analyzes and discusses the problem of reliance on expert and machine evidence, including Artificial Intelligence output, from a decision-analytic point of view. Machine evidence is broadly understood here as the result of computational approaches, with or without a human-in-the-loop, applied to the analysis and the assessment of the probative value of forensic traces such as fingermarks. We treat reliance as a personal decision for the factfinder; specifically, we define it as a function of the congruence between expert output in a given case and ground truth, combined with the decision-maker's preferences among accurate and inaccurate decision outcomes. The originality of this analysis lies in its divergence from mainstream approaches that rely on standard, aggregate performance metrics for expert and AI systems, such as aggregate accuracy rates, as the defining criteria for reliance. Using fingermark analysis as an example, we show that our decision-theoretic criterion for the reliance on expert and machine output has a dual advantage. On the one hand, it focuses on what is really at stake in reliance on such output and, on the other hand, it has the ability to assist the decision-maker with the fundamentally personal problem of deciding to rely. In essence, our account represents a model- and coherence-based analysis of the practical questions and justificatory burden encountered by anyone required to deal with computational output in forensic science contexts. Our account provides a normative decision structure that is a reference point against which intuitive viewpoints regarding reliance can be compared, which complements standard and essentially data-centered assessment criteria. We argue that these considerations, although primarily a theoretical contribution, are fundamental to the discourses on how to use algorithmic output in areas such as fingerprint analysis.

Keywords: machine evidence; AI output; normative decision structures; decision theory; fingerprints.

1. Introduction

With the increasing availability of specialized data, more and more decisions are being made that rely not only on direct observation but also on information provided by some intermediary. These intermediaries, which can be machines or humans, or systems consisting of a machine–human team, take in selected measurements, process them, and provide output in the form of a recommendation. The use of such information, which we collectively refer to as "expert output," has affected all areas of life, including the legal process, where it has reached unprecedented levels of sophistication and pervasiveness.[1] Forensic geneticists, for instance, use probabilistic genotyping systems to analyze complex DNA mixtures to help consumers of expert evidence to deal with questions such as whether or not a particular individual is a contributor to a recovered

[1] It manifests itself in what is called "machine evidence" (Roth 2016, 2017; Nunn 2019–2020) or, more generally, "AI output" (e.g. Lau and Biedermann 2020), as a special type of expert evidence.

Received: 4 August 2022. **Accepted:** 17 May 2024

DNA trace (Buckleton et al. 2019; Coble and Bright 2019).[2] Forensic fingerprint examiners use computer software to perform computations of the probative value of observed similarities and differences when comparing a fingermark to a reference print of a person of interest (POI) (e.g. Swofford et al. 2018). Similarly, facial image comparison analysts use specialized software to capture facial features, derive comparison scores, and assess the probative value of such scores (e.g. Jacquet and Champod 2020). Besides expert evidence, various stages in the legal process ranging from policing and sentencing to correction and parole have seen the development and introduction of data-driven "risk-assessment" systems to assess the probability of adverse outcomes such as recidivism (e.g. Garrett and Monahan 2019). The use of expert output is likely to further expand in the future. A special section on this topic in this journal, focusing on fingerprints, is therefore both timely and valuable for the discipline.

The use of humans and machines as intermediaries to generate expert output from data raises a number of questions and concerns (e.g. Roth 2016, 2017; Swofford and Champod 2021). The most prominent and lively debates concern observations of certain systems exhibiting varying performance characteristics, such as accuracy rates, in different demographic groups (Castelvecchi 2020). Such considerations underlie ongoing controversies about the suitability of expert systems[3] for practical deployment. More broadly, the use of expert systems and their outputs are at the center of regulatory developments, such as the Artificial Intelligence (AI) Act recently approved by the European Union (EU) Council.[4]

Overall, discussions around AI often focus on what, in the context of this article, can be considered as the *admissibility* of AI systems in the first place. While important, resolving the question of admissibility still leaves us with the practical question of what an individual consumer of expert output *ought to do* with the output of a given system that has met the requisite legal standards to be given consideration. Upon encountering such expert output, the individual actor faces the question of whether or not to actually rely on that output, that is to take it into account.

While this question is central to the theme addressed by the papers in the special section of this journal, that is envisioning a future in which algorithmic approaches are a part of the practice of fingerprint examination, we will argue in this article that even when thinking broadly about how practical proceedings in such a future might look like, we can hardly dispense with a formal and analytical approach *if* we intend future practice to involve coherent reasoning and decision-making. This article offers such a formal analysis of reliance on expert and AI output and discusses the argumentative implications of this analysis. We also anticipate that, ultimately, the question of reliance is fundamental in the sense that it applies to all types of expert and AI output, and thus there is nothing inherently special about fingerprint examination that would allow us to exempt it from the fundamental aspects of reliance that we analyze and discuss.

The bottom line is that, contrary to what is sometimes thought or implicitly assumed, there is no need to make a methodological distinction between statistical/AI models for fingerprint analysis on the one hand and discourses on how to use AI output in practice on the other. The conceptual problem of how to use AI output in practice is as amenable to and in need of formal analysis as the process of generating AI output itself. Therefore, these two aspects can indeed be logically combined.

The question of reliance on experts within the common law has been described as a problem of "deference," that is "whether fact finders are to be educated by or to defer to experts" (Allen and Miller 1993: 1131). This suggests a categorical acceptance of expert evidence that goes beyond the softer form of reliance that we consider here. The purpose of our article is to approach the notion of reliance from an alternative, more analytical perspective. Specifically, we will understand reliance as the incorporation of expert output into a decision-maker's overall knowledge base, not its dominance. We analyze reliance in this sense from a decision-theoretic perspective. By interpreting reliance as a decision, we expose and formally state the logical

[2] See, for example, the case *United States v Gissantaner*, No. 19-2305 (6th Cir. 2021), for an example of the intricate debates sparked by the introduction of results of probabilistic genotyping systems at trial.

[3] Broadly speaking, AI can be seen as a more automated form of these systems.

[4] Regulation of the European Parliament and of the Council laying down harmonized rules on AI and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (AI Act).

underpinnings of reliance decisions, emphasize the inevitability of dealing with uncertainty, and argue for the need to introduce value judgments. In combination, these ingredients reveal the dimensions that standard metrics for characterizing the performance of AI systems do not account for.

This article is organized as follows. Section 2 provides a general statement of the problem of reliance on expert output, using a generic example of information provided by a forensic fingerprint examiner. In subsequent sections, we explain and delimit the scope of our analysis, define the decision-theoretic model, and analyze its properties. The discussion and conclusions are presented in Section 3.

Methodologically, our development is based on normative decision theory (Baron 2008, 2012). At times, we use graphical models, that is influence diagrams (Kjærulff and Madsen 2008; Taroni et al. 2014), to illustrate the structural assumptions of our model and to keep track of formulaic results. The framework presented in this article uses probability as a measure of uncertainty (Lindley 1987) and loss as a measure of the undesirability of decision consequences. Our goal is to develop a rational decision structure that can be justified independently of how individuals naturally make decisions. This is in contrast to the purely observational and descriptive studies of decision-making that are commonly reported in the scientific literature. Although an important area of research, we do not consider intuitive attitudes toward formal approaches to reasoning and decision-making.[5]

2. Decision-analytic account of reliance

2.1 General statement of the problem

Consider a situation where a decision (action) needs to be taken. The decision-maker may need or benefit from assistance in understanding the available information. The decision-maker therefore needs to consider whether to rely on a source that provides him or her with an output, denoted here as E, processing the input information into a more usable form.[6] The output E can come from a human expert, a machine (including computers, instruments, etc.), or a combination of the two, such as a machine-assisted human expert.[7] In the most general sense, the output E takes the form of a recommendation about a hypothesis H, called a proposition here, based on some sort of input information or input measurement. The main example of E that we will use in this article is the report of a forensic fingerprint examiner or even a machine that observed similarities and differences between the features of a recovered fingermark of an unknown source and the fingerprint of a POI provide moderate (or some other degree of) support for the proposition that the POI, rather than an unknown person, is the source of the recovered fingermark.

However, E can also be any form of expert or machine output, including:

- a report by a forensic geneticist that the observed correspondence between the DNA profile of a recovered biological stain and that of a POI provides strong support for the proposition that the POI, rather than an unknown person, is the source of the recovered DNA;
- a report by a digital forensic examiner that the digital data support the prosecution's proposition that the phone of a POI was at the scene rather than the defense's proposition that the phone was at the home address (propositions adapted from Tart 2020);
- a report by an intelligence analyst according to which two (or more) traces (e.g. shoe marks, facial images, DNA, fibers, digital traces, bullets, etc.) seized on two (or more) distinct instances (or, locations) come from the same source, rather than from two (or more) different sources;
- a report, informed by the use of a risk assessment instrument that assigns a risk score based on certain factors (e.g. Garrett and Monahan 2019, 2020), regarding the potential of a particular individual to re-offend in the future; and

[5] See for example, Swofford and Champod (2021) for a discussion of this topic.

[6] Note that E is more than just a report or summary of measurements, such as "the general pattern of the fingermark is a whorl," or "the knife blade is 10 cm long," but adds in some form of processing. The selection of what measurements to include or even to obtain in a report or summary is a form of processing.

[7] See for example, Dror and Mnookin (2010) on the notion of distributed cognition.

- a report about a diagnostic test, asserting support for the proposition that a given individual has a particular physiological condition rather than the proposition that the individual does not have the condition of interest (e.g. Kaye 1987).

Thus, there is nothing unique or specific to fingermark evidence that distinguishes it from other types of evidence in terms of how factfinders should decide to rely on the evidence. Note, however, that we do not focus here on what Faigman et al. (2014) called "framework evidence," that is testimony limited to general statements, such as the occurrence of particular features observed on evidential items.

The expert output E is then used by the decision-maker, along with other information, to make an *ultimate decision*, with the decision to rely on E being made on the way to the ultimate decision. Crucially, the decision-maker using the output either lacks the ability or willingness to process the input information directly into E, even if the input information is available and presented to the decision-maker. Specifically, with regard to the example of fingermark evidence, the decision-maker is someone who is not a fingerprint examiner. The decision-maker will, we may hope, carefully consider the testimony of an examiner, but will not develop the expertise necessary to conduct fingermark examinations themselves. Thus, the decision to rely on the output is made without actually replicating the work, with the expert or machine being to some extent a sort of black box, even though from the expert's point of view the process used to produce the output may actually be transparent.

In the context of our running example of forensic expert evidence, machine-assisted or not, in which a fingerprint examiner reports on the comparison of a fingermark of an unknown source (e.g. found on a surface of interest) with the reference print of a POI, we focus on the following key question: what does it mean for a recipient of expert output to *rely* on the expert output? We will argue that this question of reliance on expert output is itself a *decision problem*, and that the structural features of this decision problem can be formally stated, analyzed, and discussed from a decision-theoretic perspective.

2.2 Defining the scope of analysis

When modeling inference and decision problems, it is important to be precise not only about what exactly we intend to model, but also about what we do *not* intend to model. We therefore emphasize that our analysis concentrates only on the decision of reliance on the expert output provided in the instant case. The focus is on the reliance on E, not on the ultimate decisions. Deciding to rely upon a particular output supporting a proposition X does not mean or suggest that we take that proposition X to be true, let alone that we make the ultimate decisions. For example, we may decide to take into account an expert report supporting the proposition that the particular defendant left a mark on a murder weapon *without* deciding that the defendant is the murderer or that the defendant should be convicted. Making decisions about ultimate propositions is a separate decision problem (e.g. Kaplan 1968; Kaye 1999).

Furthermore, we do not focus on the decision to consider a particular expert output as admissible. In this discussion, we assume that the output is already admissible under the applicable criteria. Given that notions of reliability are often incorporated into definitions of admissibility, it may be tempting to conflate admissibility and reliance. However, the two are logically distinct. Just because something is reliable enough to be admitted does not mean that it is reliable enough to be relied upon in making the ultimate decision, which, as we will see, includes other decision factors. Therefore, we focus here on the decision of whether or not to take an expert output into account, once that information is permitted for consideration in the ultimate decision process.

With respect to the question of how decision-makers *ought* to use the output of an information source to revise their beliefs, we will rely on principles of standard probabilistic inference (see, e.g. Kaye 1987 for an overview of general principles), although we will not address the empirical problem of the (lack of) epistemic competence of human fact finders (e.g. Mnookin 2008). For the sake of argument, we assume that data on the overall performance of the information source are available. However, we stress that this is not a necessary requirement. Our model is flexible and can accommodate modifications described in the existing literature, in

particular the endogenous definition of reliability for partially reliable sources of information when validation or accuracy studies in the traditional sense are not available or applicable (e.g. Bovens and Hartmann 2003; Lau and Biedermann 2020). The key point here is that these other modeling approaches treat *reliability* as a *property* of the information source, whereas our analysis deals with *reliance* understood as a *decision* of the fact finder.

Another aspect, not covered in our analysis, is the evaluation and comparison of the performance of different sources of information, as well as systems and procedures that produce particular (expert) output, such as forensic value-of-evidence computations (e.g. Ramos and Gonzalez-Rodriguez 2013). These are valuable considerations for experts faced with the question of selecting one of the several available expert systems, or legal decision-makers who must decide on the admissibility of a particular proffered system output. Our analysis focuses instead on a more advanced stage in the ultimate decision process, that is, the question of reliance on particular system output that has already been deemed admissible.

2.3 Modeling the problem of reliance as a decision

The problem of reliance has two main aspects. One aspect, called inference here, is about how to use the expert output to inform our view of the competing propositions. Another aspect deals with the question of whether or not to rely on the expert output. This is a question of decision. The former aspect, inference, is extensively covered in existing literature and will only be touched on briefly here. Our focus is on the latter, the decision problem, and how inference and decision can be logically related within a coherent whole. Figure 1a shows the structure of the model described in the remainder of this article. Table 1 summarizes the node definitions. Key to our model is that inference and decision are related through expressions of preferences between accurate and inaccurate decision outcomes (i.e. a so-called loss function). In this model, the notion of accuracy refers to the congruence between the expert output E and the ground truth H. It is possible to represent this aspect, accuracy, in terms of a separate node, as shown in Fig. 1b. Under certain conditions, such a model can be shown to give the same numerical results as the more compact model (a). However, the formulaic development of the second model, (b), is more burdensome. Therefore, for ease of exposition, we continue our analysis using the model shown in Fig. 1a, recognizing that the two possible models are representations of the same decision problem at different levels of resolution.

A defining feature of our model is that expert output is available and is the starting point of our analysis. That is, in our analysis, we seek to assess the "goodness" of reliance decisions in view of what our beliefs about the propositions *would* be if the expert output were taken into account. This is in contrast to the well-known model for the "Oil Wildcatter" two-decision problem (e.g. Raiffa 1968; Shenoy 1992; Cowell et al. 1999; Kjærulff and Madsen 2008), which focuses on the decision of whether or not to seek information (i.e. conducting a seismic test) *prior* to making a principal decision (i.e. drilling for oil). See for example, Biedermann et al. (2020), Gittelson et al. (2013), and Taroni et al. (2014) for forensic and legal applications of the latter model. Note, however, that there is nothing to prevent our model from being extended to handle two-decision problems.

2.3.1 *Inference*

Our analysis begins with the problem of inference, represented by the network fragment $H \rightarrow E$. The node H has two states, representing the two propositions "The POI is the source of the fingermark" (H_1) and "An unknown person is the source of the fingermark" (H_2). The two propositions H_1 and H_2 thus capture the entirety of what can happen. The decision-maker's probabilities for the propositions H_1 and H_2 are organized in a node probability table, associated with the node H. The decision-maker's probabilities are conditioned on the entirety of knowledge and information I available at the time a decision needs to be made about reliance. We write these probabilities, prior to considering evidence E, as $\Pr(H_j|I), j = \{1,2\}$, but we do not explicitly model the information I using a separate node. We also omit I from the notation for simplicity.

The node E represents the expert output, modeled as a child variable of the node H. Expert outputs are, generally, recommendations, but come in many forms. In forensic science, such

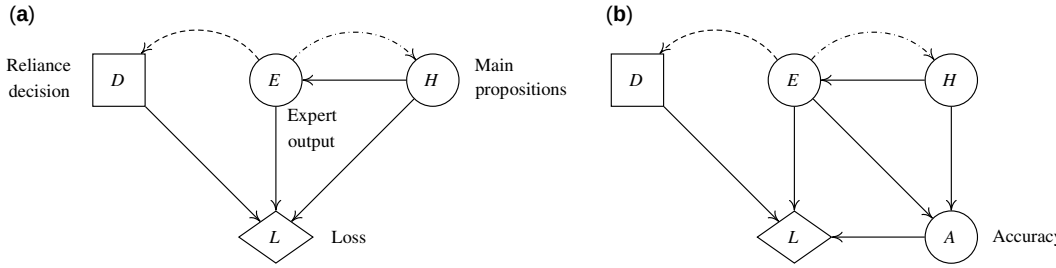

Figure 1. (a) Influence diagram for the problem of deciding whether or not to rely on expert output. The definition of the nodes *D*, *E*, *H*, and *L* is given in Table 1. The bent dashed arc is an informational link that represents the understanding that *E* is known at the time a decision is made at the node *D*. The bent dash-dotted arc represents the direction of inference when reasoning about *H* in light of the evidence *E*. (**b**) Alternative model structure, including a node *A* to monitor the congruence between the expert output *E* and the ground truth *H*.

Table 1. Definition of the nodes used in the influence diagram shown in Fig. 1a. Nodes *E* and *H* are chance (i.e. probabilistic) nodes, *D* is a decision node, and *L* is a utility (loss) node.

Node	Definition
D	Decision regarding reliance on expert output: D_1: Rely on expert. D_2: Do not rely on expert.
E	Expert output (statement or conclusion) E_i: E_1: "The findings provide support for H_1 over H_2." E_2: "The findings do not support one proposition over the other." E_3: "The findings provide support for H_2 over H_1."
H	Propositions H_j: H_1: e.g. The POI is the source of the fingermark. H_2: e.g. An unknown person is the source of the fingermark.
L	Loss function $\mathrm{L}(\cdot)$, assigning a loss value to each decision consequence $C_{ijk} = (E_i, H_j, D_k)$, for $i = \{1,2,3\}$ and $j,k = \{1,2\}$.

expert outputs vary widely in reporting style. Some experts make direct statements about disputed events, for example, that a particular fingermark came from a particular POI. Others focus on reporting only the value of their observations, that is the extent to which their observations help to discriminate between competing claims about the source of the fingermark. Our analysis here is entirely general and can accommodate any type of reporting format. For the purpose of discussion, we consider three different conclusions: E_1, the findings support H_1 over H_2; E_2, the findings do not support one proposition over the other; E_3, the findings support H_2 over H_1.[8]

Associated with the node E is a table containing conditional probabilities $\Pr(E|H)$, with $\sum_i \Pr(E_i|H) = 1$. For the time being, we will not go into the details of assigning probabilities, we simply introduce their notation, as shown in Table 2, and some general properties.[9] Note that, in general, in order for an output E_i to have probative value with respect to the propositions $H_{\{1,2\}}$, it is required that $\Pr(E_i|H_1) \neq \Pr(E_i|H_2)$. In particular, the greater the difference between $\Pr(E_i|H_1)$ and $\Pr(E_i|H_2)$, the greater the probative capacity of E_i as a piece of evidence with respect to $H_{\{1,2\}}$. Section 2.4.3 presents more detailed considerations.

[8] An alternative, but not equivalent way, of defining E is to consider the conclusions "identification" (E_1), "inconclusive" (E_2), and "exclusion" (E_3) (e.g. U.S. Department of Justice 2020). Note, however, that this is a conceptually different and in many ways problematic reporting format, for reasons widely discussed in the literature (e.g. Cole and Biedermann 2020). Nevertheless—from a practical point of view—this type of conclusion continues to be used by many practitioners (Swofford et al. 2021), so it makes sense to mention it here.

[9] In the context of medical diagnostic accuracy studies, p and $1-m-n$ are commonly equated with, respectively, the false-positive and false-negative rates, and m and $1-p-q$ are equated with, respectively, the true-positive and the true-negative rates (e.g. Shinkins et al. 2013).

Table 2. Probability table associated with node E in the model shown in Fig. 1a.

	Proposition H	
Expert output E	H_1	H_2
E_1	m	p
E_2	n	q
E_3	$1-m-n$	$1-p-q$

The bent dash-dotted edge, pointing from E to H, indicates the direction of reasoning. That is, based on learning the expert's output E, our degrees of belief in the proposition H are revised, using standard methods of probabilistic reasoning. In influence diagrams, this is operated through Bayes' theorem (e.g. Kjærulff and Madsen 2008), as will be discussed later in Section 2.4.3. Note that the two-node representation $H \rightarrow E$ is the simplest representation of an inference scheme, chosen here for the sole purpose of limiting the complexity of subsequent formulaic developments. It is possible to extend this representation by introducing additional intermediate considerations, as described elsewhere in the literature (see Section 2.5 for further details).

2.3.2 *Accuracy*

Expert output ought to be accurate. That is, the output should accord with the actual ground truth and thus give us factually correct guidance. In the context of medical diagnosis, if an individual truly has a particular disease, we expect the diagnostic test to be positive. Similarly, in the context of measuring physical quantities, accuracy refers to the distance between the true value and the value obtained during measurement.[10] In our running example, if the POI truly is the source of the fingermark (H_1), we would want the expert to report E_1, "The findings provide support for H_1 over H_2," because such a conclusion would be accurate. This is a deductive understanding of the notion of accuracy, to be distinguished from the endogenous definition of *reliability* for partially reliable sources of information modeled with a different network structure (e.g. Bovens and Hartmann 2003; Lau and Biedermann 2020). That is, in our model, outputs E_1 and E_3 are accurate whenever H_1 and H_2, respectively, are true. Note, however, that the situation is more subtle for the output of type E_2, which does not assert support for one proposition over the other. This type of output, which can be seen as a neutral stance, cannot be accurate or inaccurate in the way that E_1 and E_3 can be (Koehler 2008; Biedermann and Kotsoglou 2021; Swofford et al. 2024).

The model shown in Fig. 1a incorporates the above understanding of accuracy, although not explicitly. As mentioned earlier, the model does not contain a separate node that indicates the probability[11] of accuracy as a function of the congruence between E_i and H_j. Nevertheless, accuracy is taken into account indirectly through the node L, which is used to express preferences between the different possible decision consequences, that is accurate and inaccurate decision outcomes and situations where the expert output is E_2. Further properties of the node L are presented in Section 2.3.3.

It is important to emphasize that what is being contemplated here is not domain-wide accuracy, or the accuracy of an expert or machine *in general*, a notion widely associated with aggregate error rate measurement. Instead, what is being considered here is the accuracy of the particular expert output E that has been provided for the particular ultimate decision problem at hand.

This is a subtle and important distinction. General measures of performance, such as those provided by conventional error rates, may be a relevant preliminary consideration in deciding whether a particular type of evidence, or a particular expert in a particular field of specialization, should be admitted and heard. However, a case-specific consideration of accuracy is still

[10] This is to be distinguished from precision, which refers to the spread of repeated measurements of some quantity.

[11] Note that we are talking about the *probability* of accuracy here because, while E_i is known, H_j is not, thus introducing uncertainty about the accuracy status of E_i.

required when deciding whether or not to rely on the particular expert information for the situation at hand. Section 2.3.3 explains how this task is handled by the model shown in Fig. 1a. In summary, accuracy in the aggregate case is a preliminary consideration, but not equivalent to the accuracy of expert output in the instant case. Our model discussed here considers only the latter type of accuracy.

2.3.3 *Decision*

The part of the model that deals with the problem of *deciding* reliance needs to be understood in the context of the four nodes E, D, H, and L, which are arranged in a converging connection with E, D, and H as parent variables of node L (see Fig. 1a). Central here is node D, a decision node, with two states, D_1, relying on expert output, and D_2, not relying on expert output. The bent dashed arc pointing from E to D is an informational link, indicating that E is known at the time a decision at node D is made.[12] This accounts for the reality that there is nothing to decide about reliance as long as no expert output E is available. Note that in our model the decision is *not* whether to accept H_1 or H_2, as this would amount to a decision-theoretic account of hypothesis testing. Our analysis focuses solely on whether or not to rely on expert output regarding proposition H. That is, given the available information, reliance is about whether or not to accept the inferential impact of the information on the relevant propositions at hand.

Generally speaking, making a decision leads to a particular consequence. A consequence is defined as the combination of a decision and a particular state of nature. Decision-making under *uncertainty*, as understood here, means that we do not know what the state of nature is when we make a decision, and therefore we do not know which decision consequence will be obtained. For example, when a fact finder concludes that a fingermark was left by a particular person, the fact finder does not actually know whether that conclusion is accurate or erroneous. However, the degree to which we believe one proposition to be true rather than another can be expressed in terms of probability and be taken into account. Note that the literature sometimes distinguishes between the concepts of risk and uncertainty. According to these accounts, risk refers to situations where the decision-maker is able to specify probabilities (or chances). In turn, those accounts use the notion of uncertainty for situations in which probabilities cannot be specified and are said to be unknown or unknowable. We do not make this distinction here because probabilities are not a case of being knowable or not (de Finetti 1974; Lindley 1987). By definition, a probability is a measure of our (your, anybody's) uncertainty about the truth or otherwise of a proposition of interest. The point is that people are more or less willing to articulate their probabilities, and that there are situations in which the specification of probability is felt to be easier (e.g. the toss of a fair coin) than in others (e.g. the outcome of an election) (Lindley 1985). Thus, for the purposes of our analyses, decision-making is considered to be decision-making under uncertainty, with probabilities assumed to be ascertainable in principle.[13]

In our model, a decision about reliance is made given knowledge of the type of expert output E_i under consideration, together with expressions of (1) preferences among the possible consequences of reliance (and non-reliance) on that output and (2) uncertainty about the ground truth state H. In this framework, decisions thus lead to well-defined *consequences*. More formally, a decision consequence C is defined here as the combination of an expert output E_i, a ground truth state H_j and a decision D_k, denoted $C_{ijk} = (E_i, H_j, D_k)$, for $i = \{1,2,3\}$ and $j,k = \{1,2\}$. We do not model decision consequences explicitly here, using a separate node C.[14] Instead, we model the decision-maker's preferences among decision consequences directly, using a utility node L. For this discussion, we frame preferences between decision consequences in terms of

[12] As noted in Section 2.3, this is a major difference with respect to other models described in the literature, which focus on the decision whether or not to acquire (additional) information *before* making another decision. In our work, the situation is different: information is already available and the question (faced by the fact finder) is whether to rely on it.

[13] As an aside, it should be noted that the reader is free to reject the specification of uncertainties in terms of probabilities. However, as we will show in later parts of this article (Section 2.4.1), such a position can severely compromise decision-making procedures.

[14] See Biedermann et al. (2020) for an example of an explicit representation of decision consequences in terms of a distinct node C. It can be shown that omitting such a node C does not affect the result of the decision-theoretic computations to determine optimal decisions.

undesirability, using the concept of loss. Thus, we assign a loss $\mathrm{L}(\cdot)$ to each consequence $C_{ijk} = (E_i, H_j, D_k)$. Structurally, this view is expressed by the arcs pointing from nodes E, D, and H to node L.

What remains to be done is to specify the loss function $\mathrm{L}(\cdot)$. The specification of loss functions is an intricate topic. Arguments are needed to rationalize and justify particular loss functions. Start by considering what is called a 0 – 1 non-negative loss function.[15] This function assigns the loss value 0 to the best consequence(s), and the value 1 to the worst consequence(s). Using the value 0 for the most preferable consequence(s) expresses the idea that nothing is lost by obtaining the consequence of interest, as no better outcome could have been obtained. For our running example, we specify the following losses (see Table 3 for a summary):

- For findings E_1 that assert support of H_1 over H_2:
  - $\mathrm{L}(E_1, H_1, D_1) = \mathrm{L}(E_1, H_2, D_2) = 0$: Reliance ($D_1$) in the case of accurate output (i.e. E_1 and H_1 hold), and non-reliance (D_2) in the case of non-accurate output (i.e. E_1 and H_2 hold), are the most desirable outcomes, and thus imply zero loss.
  - $\mathrm{L}(E_1, H_2, D_1) = \ell_1 = 1$: Relying on inaccurately incriminating output represents the worst consequence. Therefore, we assign the loss value 1.
  - $\mathrm{L}(E_1, H_1, D_2) = \ell_2$, for $0 < \ell_2 \leq 1$: Not relying (D_2) on accurate output (i.e. E_1 and H_1 hold) is suboptimal, hence requires us to assign a loss different from zero. The question is whether this consequence—the failure to rely on accurate output—is as undesirable as relying on inaccurately incriminating output, that is (E_1, H_2, D_1). For shortness of notation, we will occasionally express the loss for (E_1, H_1, D_2) as ℓ_2.
- For findings E_3 that assert support of H_2 over H_1:
  - $\mathrm{L}(E_3, H_2, D_1) = \mathrm{L}(E_3, H_1, D_2) = 0$: Analogous to the assignments above, we assign a loss of zero to reliance (D_1) in the case of accurate output (i.e. E_3 and H_2 hold), and to non-reliance (D_2) in the case of non-accurate output (i.e. E_3 and H_1 hold), because these are two desirable outcomes.
  - $\mathrm{L}(E_3, H_1, D_1) = \ell_3$, for $0 < \ell_3 \leq 1$: Relying (D_1) on inaccurately exculpatory expert output (i.e. E_3 and H_1 hold) is suboptimal, that is to some extent undesirable. The question is whether this is as undesirable as failing to rely on accurately exculpatory expert output, (E_3, H_2, D_2), to which the loss value 1 is assigned. Again, for shortness of notation, we will designate the loss for (E_3, H_1, D_1) by ℓ_3.
  - $\mathrm{L}(E_3, H_2, D_2) = \ell_4 = 1$: Not relying ($D_2$) on accurately exculpatory expert output (i.e. E_3 and H_2 hold) is considered a highly undesirable result. This is analogous to relying on inaccurately incriminating output, defined above. A loss value of 1 is therefore assigned.
- For findings E_2 that do not assert support of one proposition over the other: Recall from Section 2.3.2 that output of type E_2 cannot be accurate or inaccurate in the sense that E_1 and E_3 can be. Yet, the question the decision-maker needs to ask is how the two types of losses $\mathrm{L}(E_2, H_j, D_1)$ and $\mathrm{L}(E_2, H_j, D_2)$, for $j = 1, 2$, compare to one another. Recall that, ultimately, we do not alter our belief about H in the case of an output of type E_2: if we decide D_2, not to rely on E_2, this means that we discard E_2 altogether; if we decide D_1, to rely on E_2, we do not alter our beliefs either because E_2 does not assert support for one proposition over the other.[16] However, in our view, the latter course of action would represent a greater loss, because we have the burden (or cost) of reliance on an item of information from which we derive no inferential guidance. So let $\mathrm{L}(E_2, H_j, D_1) > \mathrm{L}(E_2, H_j, D_2)$, $j = 1, 2$. For the purpose of further analysis, we set $\mathrm{L}(E_2, H_j, D_2) = 0$ and $\mathrm{L}(E_2, H_j, D_1) = \gamma$, for $0 < \gamma \leq 1$ and $j = 1, 2$.

[15] The 0 – 1 scale is often chosen because there are well-established procedures for eliciting value judgments with this scale (see e.g. Lindley 1985; von Winterfeldt and Edwards 1986).

[16] Note that this requires us to assume, following the notation introduced in Section 2.3.1, that $n = q$. This assumption may not exactly overlap with the empirical observation made in some areas of application, such as the examination of striated marks on fired bullets, that so-called "inconclusive" conclusions have different rates of occurrence under the two possible ground truth states (e.g. Hofmann et al. 2020). It should be noted, however, that such data are, at best, informative but not prescriptive for probability assignment. See also Section 2.4.3 for further discussion.

Table 3. Loss matrix for the problem of reliance on expert output. For each combination of an expert output E_i, a ground truth state H_j, and a decision D_k, a loss value is assigned on a 0–1 scale.

Loss		Output E					
		E_1		E_2		E_3	
		Proposition H					
		H_1	H_2	H_1	H_2	H_1	H_2
Decision D	D_1	0	$\ell_1 = 1$	γ	γ	ℓ_3	0
	D_2	ℓ_2	0	0	0	0	$\ell_4 = 1$

We reiterate that the loss values specified above can be broadly understood as expressions of regret related to the consequences of reliance decisions and the congruence between expert output and ground truth. We reiterate that they should not be confused with losses resulting from ultimate decisions, as ultimate decision-making may take into account other factors unrelated to belief in the facts of the situation at hand.

2.4 Analysis of model properties

The model described in Section 2.3 provides a static representation of the problem of reliance. The model clarifies the variables that are thought to capture the essential elements of the problem at hand, along with the relationships that are assumed to hold between the variables. Note that the presence of an arc between a pair of nodes, representing a direct influence of one variable on another, is as informative as the absence of an arc. For example, besides an informational link (Section 2.3.3), there is *no* straight, plain edge pointing from node E to node D. This means that expert output does not predicate or otherwise directly influence a decision about reliance.

However, this descriptive account does not tell us yet how to choose between the rival decisions D_1 and D_2, for any given type of expert output E. We need to look deeper into this formal framework to uncover its logical implications for deciding about reliance. In the next two sections, we consider different decision criteria and examine their properties.

2.4.1 *Deciding without probabilities?*

While it is uncontroversial that we ought to rely on accurate expert output, this is usually easier said than done. We cannot know for sure whether expert output is accurate or not, so we cannot readily determine which decision, D_1 or D_2, will minimize *actual* loss. An exception, however, is the case of expert output E_2, which affirms no support for one proposition over its alternative. As noted in Section 2.3.2, the notion of accuracy in the traditional sense is not applicable to such output. From the assigned loss function in Table 3, we can directly determine D_2 as the decision that minimizes loss, as $\mathrm{L}(D_2, E_2) < \mathrm{L}(D_1, E_2)$, regardless of the ground truth state H. Thus, there is no decisional problem of reliance with output of type E_2. In the remainder of this article, we will not consider this type of output any further.

For output of type E_1 and E_3, as mentioned in Section 2.3.2, accuracy depends on whether H_1 or H_2 is true. Thus, by definition, in deciding whether to rely on these types of outputs, we inevitably need to deal with probabilities. Yet, many view probabilities with skepticism. Some might even claim that, in their decision-making, they dispense with probabilities altogether. This raises the question of whether it is reasonably possible to ignore the inevitability of probability and conceive of a decision criterion that does not involve probability.[17] As we shall see, such a deterministic approach does not result in quality decision-making.

A well-known example of a probability-agnostic decision procedure is the minimax rule for individual decision-making. Minimax is a concept used more widely by game theorists to study situations where the decision-maker faces an active adversary. However, the problem of reliance

[17] See Biedermann et al. (2018) for a related discussion in the field of forensic identification.

does not involve a game in the traditional sense of game theory. Instead, our hypothetical decision-maker faces possible states of nature that define the accuracy of expert output.

Applying the minimax rule to individual decision-making involves the following steps. For each course of action (decision), identify the worst consequence (i.e. the one with the highest loss). Then choose the decision that minimizes the maximum loss over the different states of nature. Let us see where this leads in our running example for outputs of type E_1 and E_3.

To begin our analysis, consider a special case of the loss function $\mathrm{L}(\cdot)$, known as a symmetric loss function, by assigning $\ell_2 = 1$ and $\ell_3 = 1$. In words, this assignment means that, in the case of expert output E_1, reliance D_1 when H_2 is true (i.e. the output is non-accurate) is taken to be as undesirable as non-reliance D_2 when H_1 is true (i.e. the output is accurate). Similarly, in the case of expert output E_3, we assume that reliance D_1 when H_1 is true (i.e. the output is non-accurate) is as undesirable as non-reliance D_2 when H_2 is true (i.e. the output is accurate). See also Table 3 for a summary. With such a loss function, we immediately see that for each decision $D_{\{1,2\}}$, and each type of expert output $E_{\{1,3\}}$, we have exactly one optimal outcome (with loss 0), and one worst consequence (with loss 1). However, since the adverse consequence for both courses of action implies *the same loss*, here the maximum value 1, the minimax method does not help us in selecting one of the two rival decisions. In terms of the minimax criterion, the two rival courses of action appear equally suitable when a symmetric loss function is assumed.

Next, consider a more general and more realistic loss function with $\ell_2 < 1$ and $\ell_3 < 1$. Such a loss function is appropriate if we consider that adverse consequences of the two courses of action D_1 and D_2 are *not* equally undesirable. More specifically, for output E_1, that is asserted support of H_1 over H_2, we consider that relying on inaccurate output is worse than not relying on accurate output:

$$\underbrace{\mathrm{L}(E_1, H_2, D_1)}_{\ell_1 = 1} > \underbrace{\mathrm{L}(E_1, H_1, D_2)}_{\ell_2 < 1}.$$

For output E_3, that is asserted support for H_2 over H_1, we consider that not relying on accurate output is worse than relying on inaccurate output:

$$\underbrace{\mathrm{L}(E_3, H_2, D_2)}_{\ell_4 = 1} > \underbrace{\mathrm{L}(E_3, H_1, D_1)}_{\ell_3 < 1}.$$

With these loss assignments in mind, we can now apply the minimax procedure as follows:

- For expert output of type E_1: If we decide D_1, the worst that can happen is that H_2 is true and, hence, the output is inaccurate, $\mathrm{L}(E_1, H_2, D_1) = 1$. The worst that can happen under D_2 is not to rely on the output even though it is accurate, because H_1 is true, $\mathrm{L}(E_1, H_1, D_2) = \ell_2$. Since $\ell_2 < 1$, the minimax decision thus is D_2. Stated otherwise, the loss incurred with D_2, in the case of an adverse outcome, is *smaller* than the loss incurred with D_1 in the case of an unfavorable outcome.
- For expert output of type E_3: If we decide D_1, the worst that can happen is to rely on inaccurate output (i.e. H_1 is true), $\mathrm{L}(E_3, H_1, D_1) = \ell_3$. If we decide D_2, then the worst consequence is not to rely on accurate exculpatory output, $\mathrm{L}(E_3, H_2, D_2) = 1$. Thus, for $\ell_3 < 1$, the minimax decision is to select D_1, to rely on output E_3. Stated otherwise, the loss incurred with D_1, in the event of an adverse outcome, is *smaller* than with D_2 in the event of an adverse outcome.

Thus, under the chosen loss function, the minimax criterion tells us to *always* rely on the expert output of type E_3, and to *never* rely on the expert output of type E_1. Clearly, this looks unsatisfactory and may be perceived as unbalanced because it amounts to selectively depriving ourselves of one type of expert output (E_1), while systematically relying on another type of expert output (E_3). This would amount to systematically relying on information that asserts

support for one viewpoint (or party), and never relying on information that asserts support for the other. So, where lies the problem?

While the above result is due to the use of an asymmetric loss function, the problem does not lie in the choice of the actual numbers (assigned loss values). What is important is the ranking of the consequences. Specifically, for each type of expert output E_1 and E_3, the loss function defines two optimal consequences, a worst consequence and an intermediate (i.e. less than worst) consequence.

Rather, the problem lies in the minimax procedure, in particular the fact that it focuses solely on preferences among decision consequences. The driving consideration is the avoidance of maximum loss. By assuming that the worst will occur, the procedure advises to select the decision with the smallest loss in the worst case. This is a paranoid attitude to decision-making, which is appropriate in cases where there is an opponent who is able to force the maximum loss for any decision but which may not be applicable in many of the practical situations involving the use of expert output.

Indeed, while avoiding excessive losses can be seen as a laudable goal, it is important to recognize that this goal comes at a price. The price is never to rely on a type of expert output, here E_1, regardless of the *probability* of leading to an adverse outcome.[18] This brings us back to the consideration of probability that is explicitly avoided in the minimax account. What we can see is that if we want to maintain the decision to rely on output E_1, rather than systematically rejecting it in order to avoid any possibility of incurring maximum loss, we have to somehow accept the fact that incurring maximum loss is a possibility.[19] The question, then, is how to conceive of a decision criterion that allows us to rationalize reliance (in case of E_1), rather than non-reliance, *despite* the potential of reliance to lead to the worst consequence.[20] We return to this topic in the next section.

2.4.2 *Probabilistic decision criterion: Scoring rival decisions through expected loss*

The conclusion at the end of the previous section highlights the observation that, at times, we make certain decisions, such as reliance, despite their potential to lead to maximum loss, provided that the probability of a non-adverse outcome is (sufficiently) high. How high this probability *ought* to be in order to warrant a particular decision thus becomes a key question. In the remainder of this section, we will elaborate on this point by focusing only on the expert output of type E_1, although the logic of the analysis applies equally to the other expert outputs.[21]

Recall that the "problem" of deciding between reliance, D_1, and non-reliance, D_2, is due to the fact that the ground truth (variable H) is not known, and therefore the accuracy status of the expert output is affected by uncertainty. Indeed, if we knew whether the expert output E_1 was accurate, we would know that decision D_1 would lead to the best consequence. Conversely, if we knew that E_1 was *not* accurate, we would know that D_2, not to rely on the expert output, was the best decision. In other words, in situations of certainty about the relevant state of nature (variable H), we *could* minimize the *actual* loss.

However, this is not possible in cases of decision-making under uncertainty. When there is uncertainty as to whether H_1 or H_2 is true, and hence doubt as to the accuracy of expert output, the decision-maker can, at best, consider the *expectation* of loss, denoted EL here. More formally, the expected loss is the sum of the actual losses associated with a given decision (i.e. the decision consequences), weighted by their respective probabilities of occurrence. For example, to obtain the expected loss for decision D_1 in the case of expert output E_1, we need to (1) multiply the loss associated with D_1 if H_1 is true, $\mathrm{L}(E_1, H_1, D_1)$, by the probability of H_1, (2) multiply the loss associated with D_1 if H_2 is true, $\mathrm{L}(E_1, H_2, D_1)$, by the probability of H_2, and (3) add the

[18] Similarly, in the context of forensic inference of source, the minimax criterion leads to the advice to never identify if the adverse outcome of identification, that is a false identification, is considered to be worse than the adverse outcome of not identifying (a person or object as the source of a given trace or mark), that is a missed identification (Biedermann and Vuille 2018b; Biedermann et al. 2018).

[19] Note that this is common in many activities of daily life. We engage in certain activities, such as modes of travel, despite their potential for highly adverse consequences.

[20] Similarly, in the case of expert output of type E_3, we may wish *not* to rely on the expert output *despite* the potential of this decision to lead to a maximum loss.

[21] See Appendix 1 for comments on the expert output of type E_3.

two products resulting from steps (1) and (2). The general formula, for any decision D_k and type of expert output E_i is:

$$\mathrm{EL}(D_k|E_i) = \sum_j \mathrm{L}(E_i, H_j, D_k) \times \Pr(H_j|E_i),\ i = \{1,2,3\},\ j,k = \{1,2\}. \tag{1}$$

Recall from the loss function defined in Table 3 that $\mathrm{L}(E_1, H_1, D_1) = 0$ and $\mathrm{L}(E_1, H_2, D_1) = 1$, hence Equation (1) for decision D_1 and expert output E_1 reduces to:

$$\mathrm{EL}(D_1|E_1) = \Pr(H_2|E_1).$$

Thus, the expected loss of the reliance decision D_1 is equal to the probability that proposition H_2 is true, that is the probability that expert output E_1 is inaccurate. This is a fairly intuitive property: the expected loss of the decision to rely on expert output E_1 is a function of the probability that the ground truth state asserted by the expert output is true. More explicitly, the higher the probability of H_1 given E_1, the smaller the expected loss of relying on expert output E_1. This is because $\Pr(H_2|E_1) = 1 - \Pr(H_1|E_1)$.

Proceeding analogously for the decision D_2, nonreliance, Equation (1) in the case of expert output E_1 allows us to find:

$$\begin{aligned}\mathrm{EL}(D_2|E_1) &= \mathrm{L}(E_1, H_1, D_2) \times \Pr(H_1|E_1) \\ &= \ell_2 \times \Pr(H_1|E_1).\end{aligned}$$

In a sense, the expected loss $\mathrm{EL}(D_k|E_1)$ of a decision can be thought of as a measure of what to expect in terms of loss when making decision D_k, given knowledge of E_1. We insist that this is a case-based assessment. In particular, we make no claim about the average loss of making decisions of type D_k in a series of unrelated cases involving distinct expert outputs of type E_1.[22]

The question remains as to how the notion of EL can help in decision-making about reliance on expert output E_i. The point is that the notion of EL provides a criterion that allows us to compare rival decisions about reliance in order to guide decision-making about *reasonable* reliance. Being able to compare rival decisions allows us to rank them: the decision-maker can consider whether reliance is, so to speak, "better" than nonreliance, in terms of expected loss.[23]

From a decision-theoretic perspective, the criterion for making the reliance decision is based on the smallest expected loss. Thus, to inquire about the conditions under which decision D_1, in the case of expert output E_1, is preferable to decision D_2 means to inquire about the following relationship between the expected loss of decisions D_1 and D_2:

$$\mathrm{EL}(D_1|E_1) < \mathrm{EL}(D_2|E_1). \tag{2}$$

To examine this expression, we start by considering—as in Section 2.4.1—a symmetric loss function with $\ell_2 = 1$. This particular loss assignment means that we consider the consequences $\mathrm{L}(E_1, H_2, D_1)$, reliance in the case of nonaccurate output, and $\mathrm{L}(E_1, H_1, D_2)$, non-reliance in the case of accurate output, to be equally undesirable (see also Table 3). Under this assumption, Expression (2) says that decision D_1 has the *smaller* expected loss than (and is therefore preferable to) decision D_2 if the posterior probability of the first proposition, $\Pr(H_1|E_1)$, is *greater* than

[22] Claims about the performance of a decision rule over many cases are essentially unwarranted because it is not known what the respective proportions of ground truth are in the two relevant categories. See also Dekay (1996) for a general discussion of this observation in the context of standards of proof and legal verdicts.

[23] This formulation bears resemblance to arguments previously formulated, such as Judge Learned Hand's well-known calculus of negligence:

> Since there are occasions when every vessel will break from her moorings, and since, if she does, she becomes a menace to those about her; the owner's duty, as in other similar situations, to provide against resulting injuries is a function of three variables: (1) The probability that she will break away; (2) the gravity of the resulting injury, if she does; (3) the burden of adequate precautions. Possibly it serves to bring this notion into relief to state it in algebraic terms: if the probability be called P; the injury, L; and the burden, B; liability depends upon whether B is less than L multiplied by P: that is whether B < PL.
>
> *United States v Carroll Towing Co.*, 159 F.2d 169, 173 (2d Cir.1947).

the posterior probability of the alternative proposition, $\Pr(H_2|E_1)$. However, since the variable H is binary, the decision criterion, Expression (2), comes down to inquiring whether the probability $\Pr(H_1|E_1)$ is greater than 0.5.[24]

The result of our analysis here makes perfect sense because if the probability that the expert output E_1 is accurate (i.e. H_1 is true) is *greater* than the probability that E_1 is not accurate (i.e. H_2 is true), the assumption of a symmetric loss function implies that the decision D_1 offers the smaller probability of incurring an undesirable outcome than D_2. More generally, the greater the probability $\Pr(H_1|E_1)$, the more advantageous it becomes to decide D_1 rather than D_2. Stated otherwise, deciding D_2, not to rely on expert output, would be suboptimal as it would have a *higher* probability for an undesirable outcome, than deciding D_1. Note that this is a first advantage over the minimax procedure described in Section 2.4.1, which cannot provide guidance in the case of a symmetric loss function.

Nevertheless, the result of our analysis is rather unspectacular. The above result comes down to the recommendation that, when faced with two courses of reliance, where the best outcomes are as desirable as the worst outcomes are undesirable, we should make the decision of reliance based on the greater (smaller) probability of obtaining a desirable (undesirable) outcome. Intuitively, we already use this rationale in our daily lives.[25]

But what if the assumption of a symmetric loss function is considered unsuitable, that is the adverse outcomes of D_1 and D_2 cannot be taken to be equally undesirable? Indeed, more realistically, we might want to consider the adverse outcome of relying on expert output E_1, consequence (E_1, H_2, D_1), to be *worse* than the adverse outcome of deciding not to rely on E_1, consequence (E_1, H_1, D_2). Specifically, where E_1 is an expert's report on a fingermark comparison in criminal litigation, consequence (E_1, H_2, D_1) means erroneously leaning toward associating a POI with a fingermark that in fact comes from an unknown person. This outcome can reasonably be considered worse than consequence (E_1, H_1, D_2), failing to rely on information that correctly tends to associate the POI with the fingermark. In our notation, such a preference structure is written as

$$\mathrm{L}(E_1, H_2, D_1) > \mathrm{L}(E_1, H_1, D_2),$$

and amounts to an asymmetric loss function. Given such a loss function, we can ask under what conditions the decision D_1, to rely on E_1, is preferable to the decision D_2, not to rely on E_1. That is, we want to investigate the relationship between the expected loss of decisions D_1 and D_2, respectively, as defined by Expression (2). Let us write the expressions for the EL in full detail, invoking Equation (1), and assigning zero loss to the desirable decision consequences (E_1, H_1, D_1) and (E_1, H_2, D_2), and then rearrange the terms:

$$\begin{aligned} &\underbrace{\mathrm{L}(E_1, H_2, D_1)}_{\ell_1} \times \Pr(H_2|E_1) < \underbrace{\mathrm{L}(E_1, H_1, D_2)}_{\ell_2} \times \Pr(H_1|E_1) \\ &\Pr(H_2|E_1)/\Pr(H_1|E_1) < \ell_2/\ell_1 \\ &\Pr(H_1|E_1)/\Pr(H_2|E_1) > \ell_1/\ell_2 \end{aligned} \tag{3}$$

Expression (3) states that in the case of expert output E_1, the expected loss of decision D_1, reliance on E_1, is smaller than the expected loss of decision D_2, non-reliance on E_1, whenever the odds in favor of proposition H_1 over H_2, given expert output E_1, exceed the ratio of losses associated with adverse consequences of decisions D_1 and D_2, respectively.[26]

Expression (3) is well suited to help examine the argumentative implications of assuming an asymmetric loss function, $\ell_1 \neq \ell_2$. In particular, we can now see that the *more* the losses ℓ_1

[24] Readers familiar with applications of decision theory in the law, following Kaplan (1968), will immediately recognize the analogy between this result and the decision-theoretic account of the preponderance of the evidence standard in civil cases. For a detailed exposition of this well-known result, see also Kaye (1999).

[25] It suffices to think of examples involving matters of life and death.

[26] This result is also known in the literature on probabilistic machine learning (e.g. Murphy 2012). For a discussion of the analogy between this result and Archimedes' law of lever, see Biedermann et al. (2016, 2020).

and ℓ_2 differ, the *higher* the required odds in favor of H_1 over H_2 in order to satisfy Expression (2).[27] Stated otherwise, Expression (3) asks us to compare—or so-to-say "weigh"—relative beliefs against relative losses (of adverse decision consequences).

Let us look at some numbers to get a feel for this relationship. Let us start with a short digression back to a symmetric loss function, assuming $\ell_1 = \ell_2$. We can immediately see from Expression (3) that the required odds in favor of H_1 over H_2 must be greater than evens. This is in agreement with the result $\Pr(H_1|E_1) > 0.5$ found in the above discussion of Expression (2) and the condition under which D_1 is the optimal decision when assuming a symmetric loss function. Figure 2a provides a graphical illustration of this result: the lines representing the expected loss of decisions D_1 (solid line) and D_2 (dashed line) intersect at $\Pr(H_1|E_1) = 0.5$ (vertical dotted line).[28] That is, for $\Pr(H_1|E_1) < 0.5$, decision D_2 has the smaller expected loss, whereas for $\Pr(H_1|E_1) > 0.5$, decision D_1 has the smaller expected loss. As an example, consider a case where $\Pr(H_1|E_1) = 0.4$, highlighted in Fig. 2a with a vertical dotted line: here, $\mathrm{EL}(D_1|E_1) = 0.6$ and $\mathrm{EL}(D_2|E_1) = 0.4$, so D_2 is the decision that minimizes the expected loss.

Now suppose that ℓ_1 is *not* equal to ℓ_2, but greater, say ten (a hundred, a thousand, etc.) times greater. This means that reliance on nonaccurate output is taken to be worse than nonreliance on accurate output. A common proxy to investigate this assumption is Blackstone's formulation: "It is better that ten guilty persons escape than that one innocent suffer" (Blackstone 1769: 352). We call this a proxy here because Blackstone's formulation tends to refer to error *rates* across multiple cases (Dekay 1996; Kaye 1999) and also to ultimate decision-making about guilt, whereas our focus here is on the relative losses involved in reliance decisions and the relationship of one person with a particular crime scene artifact.[29]

Following this line of reasoning, suppose that ℓ_1 is 10 times greater than ℓ_2, that is the loss ratio is $\ell_1/\ell_2 = 1/0.1 = 10$. Expression (3) now states that D_1 is optimal, that is has a smaller expected loss than decision D_2, for odds $\Pr(H_1|E_1)/\Pr(H_2|E_1) > 10$. This is equivalent to requiring that $\Pr(H_1|E_1)$ is greater than $10/11 = 0.91$ (rounded result). Figure 2a shows this result graphically: the lines representing the expected loss of decisions D_1 and D_2 intersect at 0.91.[30]

The same result is shown in Fig. 2b. More generally, Fig. 2b shows the minimum probability $\Pr(H_1|E_1)$ required for decision D_1 to have the smaller expected loss than decision D_2, as a function of the loss ratio ℓ_1/ℓ_2. As can be seen, for loss ratios of the order of one hundred and above, the threshold probability tends to 1 to an extent that in practice becomes increasingly difficult to conceptualize and articulate in entirely nonnumerical terms. This does not render this result useless. Quite to the contrary, it retains its practical relevance. While we can easily think of loss ratios in terms of orders of magnitude (e.g. tens, hundreds, etc.), the guidance for practical thinking can be expressed in the following terms: "The larger your ratio of losses associated with adverse decision consequences, the more certain you should be that H_1 is true rather than H_2, given E_1, for D_1 to be the better decision than D_2." Furthermore, our analysis shows how reliance decisions can reveal the nature of the loss functions themselves. This allows decisions of reliance to expose the extent to which other factors may govern the ultimate decision-making as well as to compare reliance decisions over multiple decisions.

However, it would go beyond the scope of our analysis if we interpret our decision-theoretic result in a prescriptive way (Biedermann et al. 2020), not least because our development is based on a model that captures only some but not all dimensions of the decision problem at hand. Nevertheless, the aspects we do cover—such as accuracy[31] and (un-)desirability of decision

[27] Again, this is not a new finding for legal scholars. See, for example, the discussion in Friedman (2018: 1590): "Suppose that Option One has far worse consequences if wrong than does Option Two. Then a sensible decision-maker will choose Option One rather than Option Two only if she has a high degree of confidence that Option One rather than Option Two is correct, or, put another way, only if she thinks Option One is far more probable than Option Two."

[28] Note that $\mathrm{EL}(D_2|E_1) = \Pr(H_1|E_1)$ because ℓ_2 is set to 1.

[29] So, to be clear, what we mean here is the idea that reliance on nonaccurate output E_1 is more serious, and therefore more undesirable than nonreliance on accurate output E_1. Blackstone's dictum is merely a device to help us think about and articulate orders of magnitude of relative loss in the individual case.

[30] Conversely, if we consider that reliance on inaccurate output is *less* undesirable than nonreliance on accurate information, then the lower the requirement of $\Pr(H_1|E_1)$ for reliance. This includes cases where the probability $\Pr(H_1|E_1)$ is actually *smaller* than the probability $\Pr(H_2|E_1)$, that is when $\Pr(H_1|E_1) < 0.5$.

[31] Note that, as mentioned in Section 2.3.2, we define accuracy as the congruence between the expert output and ground truth.

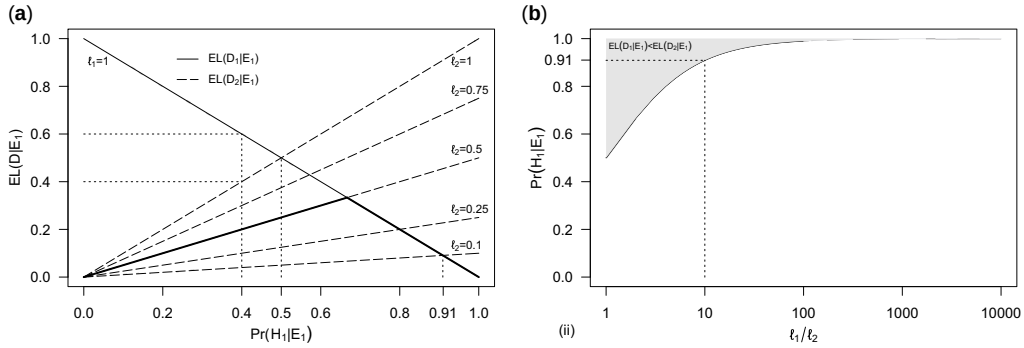

Figure 2. (a) Expected loss of decisions D_1 (relying on expert output E_1; solid line) and D_2 (nonreliance; dashed line) as a function of $\Pr(H_1|E_1)$ using a $0-1_{\{.\}}$ loss function with $l_1=1$ (loss of relying on inaccurate output E_1) and $l_2 = \{0.1, 0.25, 0.5, 0.75, 1\}$ (loss of nonreliance on accurate output E_1). The bold line highlights the optimal decision, as a function of $\Pr(H_1|E_1)$, for a case in which ℓ_1 is twice as great as ℓ_2. (**b**) Minimum probability $\Pr(H_1|E_1)$ necessary for the expected loss of decision D_1 to be smaller than the expected loss of decision D_2, as a function of the loss ratio ℓ_1/ℓ_2. The gray-shaded area shows pairs of values for the loss ratio and probability $\Pr(H_1|E_1)$ for which D_1 minimizes expected loss.

consequences—are fundamental. Furthermore, Fig. 1a represents the relationships between the different variables in a way that corresponds to our understanding of the decision problem at hand. For example, the converging connection at node L represents our understanding that we are focused on the relative value of decision consequences, given an observed value for E and different possible ground truth states H.

Our development, based on expected loss, thus provides *one* way of quantifying the "goodness" of a decision, taking into account a selected number of variables (Biedermann et al. 2018). We recognize and accept that, beyond the limited number of aspects covered by our formal development, there may be yet other considerations that govern practical decision-making. What we emphasize here is that for decision-makers who are concerned about the losses associated with uncertain decision consequences, there is a way of formally expressing relevant considerations in terms of "weighing" the losses associated with adverse decision consequences against the probability of their occurrence.

2.4.3 *Probabilities for competing propositions*

Our analysis so far shows that a decision to rely on expert output depends critically on our degree of belief in the truth state of the main proposition H. The degree of belief we have in the truth or otherwise of H is conditioned on the type of expert output (i.e. state of the variable E). More generally, as noted in Section 2.3.1, the degree of belief that we have in the truth or falsity of H is based on the totality of our knowledge and information available at the time that a decision about reliance needs to be made. However, in our account, we have been rather agnostic about particular values for these probabilities. Rather, we have made general statements about the constraints on the *orders of magnitude* of beliefs compared to the relative losses of adverse outcomes, in order to warrant particular decisions. Below we explore some further technicalities.

Recall the graphical specification of our model (Fig. 1a). The graphical structure implies two aspects that are essential for the probabilities of target propositions. The first is the temporal ordering between the node E, representing the expert output, and the decision node D. As noted in Section 2.3.3, this ordering reflects the understanding that some expert output E must be available *prior* to a decision at node D.[32] The second aspect concerns the relationship between the variables E and H, since expert output E is information that is potentially[33] pertinent to the variable H. Thus, since a decision analysis requires some expert output E as a starting point, the

[32] Thus, the question here is not whether or not to acquire evidence E. Rather, evidence E is available and the question is whether or not to rely on it.

[33] Note that whether or not E is pertinent to H depends on the conditional probabilities assigned in the node table of E.

model structure implies that the probability of H, which is used in the computation of expected loss, is *conditioned* on E, that is $\Pr(H|E)$. This amounts to a subtle point in hypothetical reasoning: when we make a decision about whether to rely on some expert output E, we should do so in the light of our degree of belief in the proposition H *conditioned* on the expert output under consideration. This may seem circular, but it is needed in order to properly compute expected loss. After all, in order to properly score rival decisions regarding reliance, the computations of expected loss of these rival decisions are necessarily based on what our beliefs *would* be if we took into account the expert output at hand.

The question that remains is how to arrive at $\Pr(H|E)$. Conceptually, assigning a value to $\Pr(H|E)$ is itself a decision, and the constraints on that decision can be stated in decision-theoretic terms, using notions such as scoring rules, originally developed in the context of evaluating and comparing the performance of forecasters (Brier 1950).[34] From a Bayesian epistemological point of view, $\Pr(H|E)$ is the result of probabilistic belief updating based on a partially reliable source of information (e.g. Bovens and Hartmann 2003). In practice, most, if not all, sources of information are only partially reliable to some extent. That is, the information—here expert output E—occurs not only when a given proposition of interest is true, but also when that proposition is *not* true. A fully reliable source of information would, among other characteristics, never provide output that asserts support for a proposition that is not true, or in other words, has a false-positive rate of zero. In practice, this is almost never the case. We will therefore use less hypothetical values.

Recall the notation introduced in Section 2.3.1 for $\Pr(E_i|H_1)$ and $\Pr(E_i|H_2)$, for $i=\{1, 2, 3\}$ (Table 2). Considering expert output E_1, asserted support for H_1 over H_2, to be informative beyond what is already supported by I means to assume the following:

$$m > p > 0.$$

Thus, the value of information of E_1 with respect to H_1 and H_2 is the likelihood ratio m/p, and takes values greater than 1.[35] We do not consider cases where $m < p$, because this would mean to take the expert output E_1, asserted support H_1 over H_2, as weakening our preexisting belief in H_1. To make this clearer, this would be equivalent to a case where a weather forecaster's announcement of rain would make us less persuaded of rain, even when we think that it would rain based on what we can see from the window.

For the remainder of this section, our discussion will focus only on expert output of type E_1. Readers interested in evaluating the expert output of type E_2 will need to assign values to n and q. The remaining probabilities, $\Pr(E_3|H_1)$ and $\Pr(E_3|H_2)$, which characterize the probative value of expert output of type E_3, are set once we define values for m and p, and n and q, respectively (see also Table 2).

To arrive at $\Pr(H|E)$, we disentangle the probative value of E from the initial belief statement about H by rewriting the left-hand side of Expression (3), using E_1 as a running example:

$$\underbrace{[\Pr(E_1|H_1)/\Pr(E_1|H_2)]}_{\text{likelihood ratio } (m/p)} \times \underbrace{[\Pr(H_1)/\Pr(H_2)]}_{\text{prior odds}}. \tag{4}$$

The term on the left is the likelihood ratio of E_1 with respect to H_1 and H_2. The term on the right is the (prior) odds of H_1 against H_2. Expression (4) says that whatever our odds on H_1 versus H_2, learning that evidence E_1 is available will multiply those odds by the factor given by the likelihood ratio.

It is important at this point to include some comments about the likelihood ratio m/p. The ratio m/p is the value that the decision-maker assigns to expert output E_1, indicative of the importance of E_1 to the overall decision-making if E_1 were to be relied upon. This ratio m/p must be

[34] See Biedermann et al. (2013, 2017) for discussions of the use of scoring rules for probability assignment in forensic science.

[35] Note that the model considered here is flexible enough to allow a decision maker to directly adopt a reported likelihood ratio if the expert output E_1 is provided in such a format.

distinguished from any statement about the probability of H_1 and H_2, including likelihood ratios, reported by the expert. Recall that, following the definition given in Section 2.3.1, expert output E_1 can refer to either a likelihood ratio with respect to H_1 and H_2, or a direct and categorical assertion regarding the truth or falsity of H_1. Conversely, the likelihoods m and p refer to the occurrence of a report of type E_1 given H_1 and H_2, respectively, *regardless* of the actual strength of support asserted by E_1. This means that even if an expert reports that the findings support H_1 over H_2 in terms of a likelihood ratio of, say, a million, or makes a categorical assertion that the POI is the source of a given fingermark, the decision about reliance requires asking "what is the probability that *this* expert would assert E_1 *in this particular case* if $H_{\{1,2\}}$ were true?". This can barely lead to a likelihood ratio of a million or more as this would require the assumption that, for example, the sensitivity m is 1 and the probability of a false-positive report p is as small as one in a million.[36] Few if any forensic disciplines or individual scientists can provide empirical support for such figures.[37]

However, even if we had such relevant data, it would still not be sufficient to obtain values for m and p. The reason for this is that data from accuracy or validation studies only provide information about the examiner's diagnostic performance *in the aggregate case*, which should not be confused with, and is not informative about, the selectivity of the features actually observed in the case at hand. By definition, a rate, proportion, or relative frequency of cases (outcomes) of a certain type is not equivalent—though related—to the probability of a single event (Lindley 2006), a distinction that is often misunderstood in the forensic science literature.[38] For decisions of reliance in routine situations, where consequences are minor, we may equate m and p with relative frequencies. For more serious decisions, such as those involving forensic science, the use of m and p from standard validation studies is at best a crude proxy that may need to be adjusted according to the circumstances of the particular case.

Notwithstanding these intricacies, data from accuracy or validation studies, including in particular proficiency testing of examiners, provide a relevant starting point for decision-makers. They can use it to provide an anchor, as argued by Koehler (2008). Champod et al. (2020) provide an illustrative example of how proficiency test data can be obtained at the individual examiner level and used to pragmatically assess the trustworthiness of a fingerprint examiner. However, when using data from validation studies, it is recommended to ensure that the data collected relate to test items and experimental conditions that reflect casework conditions, for example, in terms of the level of difficulty imposed by the quality of the items submitted for examination (i.e. degradation, pollution, etc.). Imwinkelried has referred to this as "the general notion of range of validation" (Imwinkelried 2020). For a similar viewpoint in the context of forensic voice comparison and the results of likelihood ratio computation procedures, see also Morrison et al. (2021).

However, one type of data (source) alone cannot fully resolve the assessment of expert or method performance. There are a number of ways in which relevant information can be obtained. For example, in the field of proficiency testing, at least two types of tests should be distinguished (Koehler 2017). One, more general, focuses on the ability of examiners to follow standard examination procedures. The other, more focused on operational aspects, is concerned with performance under varying casework conditions. Another dimension that decision-makers should consider is whether or not the testing was conducted in a blind manner (Mejia et al. 2020). The purpose of blind testing is to prevent examiners from approaching proficiency testing in a different, potentially more cautious or prudent, manner than they would approach regular casework, thereby compromising the validity of the resulting data. More generally, since all

[36] See Thompson et al. (2003) for a general likelihood ratio development that includes a probability for the event of a falsely positive expert conclusion, applicable to the evaluation of results of comparative forensic examinations (e.g. comparison of DNA profiles). This development makes a logical distinction between an actual correspondence between analytical features, on the one hand, and a correspondence between compared features as reported by a scientist, on the other. The former takes into account the rarity of the features in the relevant population, whereas the latter takes into account the probability of (human) error. See also Section 2.5 for further discussion and Aitken et al. (2020, section 6.1.8.4) for an overview of the impact of the potential for error on the value of evidence.

[37] See, for example, the PCAST Report (PCAST 2016) for a review of pre-2016 studies on the performance of fingerprint examiners.

[38] For discussions, see, for example, Taroni et al. (2016) and Biedermann and Vuille (2018a).

currently available data are imperfect in some way, it is recommended that data be used with caution.

It is worth noting that our model is not prescriptive about the data to be used to inform probabilities, nor about the methods of assigning probabilities. Furthermore, the model is flexible in the sense that it allows, but does not require, for example, a reported likelihood ratio, expert output E_1, to be equated with the decision-maker's likelihood ratio m/p. However, this can be problematic, especially when the reported likelihood ratios are in the billions, trillions, and beyond. Indeed, what many such likelihood ratios take into account is *only* the assessed rarity of occurrence of analytical traits. However, the target is not the abstract occurrence of analytical traits in examined items, but the real-world event of a process which has at its core an expert (or machine) providing particular output.[39] This again emphasizes the importance of information about the performance of the individual examiner and/or method in case-relevant circumstances, as noted above.

We now return to Expression (4). Readers may find the combination of m/p with prior odds by multiplication, followed by a comparison with the loss ratio ℓ_1/ℓ_2, defined by Expression (3), difficult to digest. However, it is possible to make the combination of evidential value (m/p) and prior beliefs, and the subsequent comparison with losses, more intuitive by using the logarithm (Good, 1950). We therefore rewrite Expression (3), replacing the left-hand side with Expression (4), to obtain:

$$\log_{10}\left[\frac{m}{p}\right] + \log_{10}\left[\frac{\Pr(H_1)}{\Pr(H_2)}\right] > \log_{10}\left[\frac{\ell_1}{\ell_2}\right]. \tag{5}$$

The $\log_{10}$ transformation makes the combination of likelihood ratios and odds additive. Readers may notice that working with the logarithm introduces symmetry. For example, the odds of H_1 versus H_2 of 10 (to 1), that is a $\log_{10}$ of 1, correspond to a segment on a number line between 0 and 1. Conversely, odds on H_2 against H_1 of 10 (to 1), that is a $\log_{10}$ of –1, correspond to a segment on a number line of the same length, but going from –1 to 0.

The reformulated decision criterion in Expression (5) states that the decision-maker ought to compare the entirety of evidence available at the time a decision needs to be made—expert output E_1 combined with any previous knowledge and data—with the magnitude of the ratio of losses associated with adverse decision consequences. Table 4 provides a few examples. Suppose the decision-maker receives expert output E_1 to which he assigns a likelihood ratio of 1000, that is a $\log_{10}$ of 3. Suppose further that the prior odds are even (1 : 1), corresponding to a $\log_{10}$ of 0. The combination of likelihood ratio and prior odds in terms of their $\log_{10}$ gives 1000, that is a $\log_{10}$ of 3. This is the limiting value that, according to Expression (5), the ratio of losses ℓ_1/ℓ_2 must *not* exceed for decision D_1 (reliance on expert output E_1) to be preferable to decision D_2 (not relying on expert output E_1). For $\log_{10}$ prior odds other than zero, for example, $-2, -1, 1, 2$, the limiting loss ratio changes accordingly (see Table 4).

As noted previously, we are not asking the reader to pin down specific numbers but to think in terms of general orders of magnitude. The broader conclusion here is that if ℓ_1, the loss of the adverse outcome of reliance, is, say, x times greater than ℓ_2, the loss of the adverse outcome of non-reliance, then coherence requires that we ought to be at least x times more certain that H_1 is true than H_2, given the entirety of available knowledge and data, including expert output E_1.

To some extent, this finding runs counter to common understandings in debates about expert evidence. Mainstream arguments regarding reliance are often narrowly framed in terms of aggregate performance measures, such as rates of false positives and false negatives in suitably designed[40] validation studies, which ought to be satisfied. However, there is no agreed, specific performance characteristic or threshold that must be met for a particular expert output to be considered reliable. This is rather unsurprising, since attempting to define a generic criterion for reliance on expert output would mean to dissociate expert evidence from the instant case, and

[39] As Zabell concisely noted, "the siren-like attraction of astronomically small probabilities can often blind one to their practical limits. In the end the value of forensic or any other type evidence is totally dependent on the reliability and validity of the process by which it is generated" (Zabell 2012).

[40] A suitably designed study is one that reflects the conditions of the case under consideration.

Table 4. Limits that the ratio of losses ℓ_1/ℓ_2 must *not* exceed in order for expert output E_1 (i.e. asserted support for H_1 over H_2), in the form of a likelihood ratio m/n of 10^3, combined with exemplary prior odds (PO, for H_1 over H_2), to make decision D_1 (reliance on expert output E_1) preferable to decision D_2 (not relying on expert output E_1).

m/n; $\log_{10}(m/n)$	PO	$\log_{10}$(PO)	Limiting value ℓ_1/ℓ_2	$\log_{10}(\ell_1/\ell_2)$
1000; 3	1:100	–2	10	1
	1:10	–1	100	2
	1:1	0	1,000	3
	10:1	1	10,000	4
	100:1	2	100,000	5

thus from considerations of the inconvenience of adverse consequences of reliance in the instant case. Our account avoids this impasse by embedding probative value in a coherent weighing procedure along with two other essential ingredients: prior information and relative losses associated with the decisions of reliance and nonreliance on expert output.

Figure 3 provides a graphical illustration. It shows plots of examples of likelihood ratio values $m/p = \{1, 10, 100, 1000, 10000\}$ that indicate the prior probability $\Pr(H_1)$ that must be exceeded in order to result—when combined with expert output E_1—in posterior odds that exceed the loss ratio ℓ_1/ℓ_2 and thus make decision D_1 (reliance on expert output E_1) preferable to decision D_2 (not relying on E_1), in the sense defined by Expression (3). For example, suppose the decision-maker's loss ratio ℓ_1/ℓ_2 is 10^4 and there is expert output E_1 to which a likelihood ratio m/p of 10^3 is assigned. In such a case, decision D_1, that is relying on expert output E_1, is optimal only if the prior probability $\Pr(H_1)$ is greater than 0.91, that is the prior odds $\Pr(H_1)/\Pr(H_2)$ are greater than $10:1$. If the decision-maker's prior probability is lower, then decision D_1 is preferable to D_2—that is optimal in the sense of Expression (3)—only if the loss ratio is lowered accordingly. Conversely, if the assigned likelihood ratio m/p is 10^4 a prior probability greater than 0.5 is sufficient to ensure that D_1 is the optimal decision. Figure 3 illustrates these two cases with the horizontal dashed lines at 0.91 and 0.5, respectively.

Note that this account does not exclude the limiting case of reliance on expert output E_1 to which the decision-maker assigns no value, that is a likelihood ratio m/p of 1. This is expert output for which we take the sensitivity to be equal to the probability of a false positive. In other words, the decision-maker does not consider the expert output to add any value over and above the other evidence already considered. In such a case, the prior $\Pr(H_1)$ is unaffected by the expert output and Expression (3) amounts to a direct comparison of the prior odds against the loss ratio ℓ_1/ℓ_2. Figure 3 illustrates this for a case where the loss ratio is 10 (vertical dashed line). Here, prior odds greater than $10:1$, that is $\Pr(H_1) > 0.91$, are required for decision D_1 to be preferable to decision D_2. It may seem strange that D_1 can be optimal at all, even though E_1 appears to be considered uninformative.

There are two ways to rationalize this. First, it is important to remember that the comparison of rival decisions in terms of Expression (3) is based on a loss function for outcomes defined in terms of the variable H and the type of expert output E. However, the procedure is flexible regarding the decision-maker's assigned likelihood ratio m/p. Stated otherwise, while the question of whether or not to rely on a given output E_1 is a function of the ground truth H, there is nothing in the procedure that prevents the decision-maker from considering the output to have no probative value. Second, and related to the previous point, the probability of H_1, that is the proposition toward which the expert output tends, may already be sufficiently high (e.g. $\Pr(H_1) = 0.91$ as noted above) based on *other* available evidence, thus warranting decision D_1 even though the specific expert's output is assigned no value. In such a case, reliance D_1 would mean to acknowledge the direction of inference conveyed by the expert even though we consider the expert in fact adds no information to what is already known. Incidentally, this is what distinguishes $(m/p) = 1$ assigned to the *assertive* expert output E_1 from the expert output E_2 which asserts *no* support for one proposition over the other.

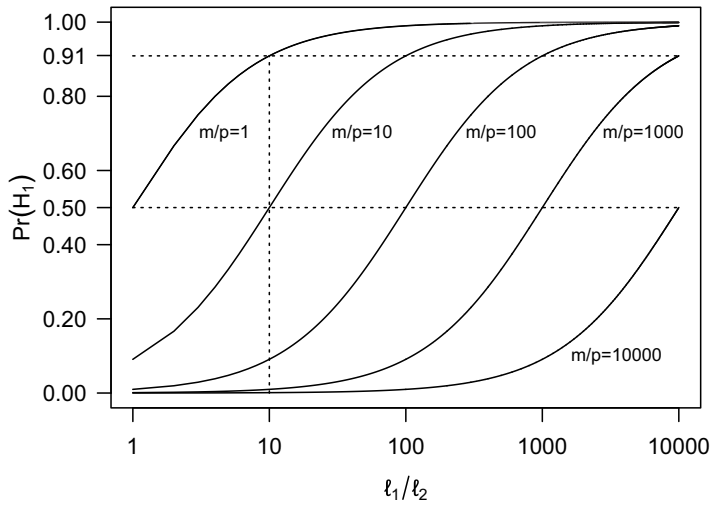

Figure 3. Plot of curves representing examples of likelihood ratio values m/p for expert output E_1, indicating the prior probability $\Pr(H_1)$ (y-axis) that must be exceeded in order for the posterior odds $\Pr(H_1|E_1)/\Pr(H_2|E_1)$ to be greater than the loss ratio ℓ_1/ℓ_2 (x-axis) according to Expression (3), making decision D_1 preferable to decision D_2. The dashed lines highlight specific cases discussed in the text.

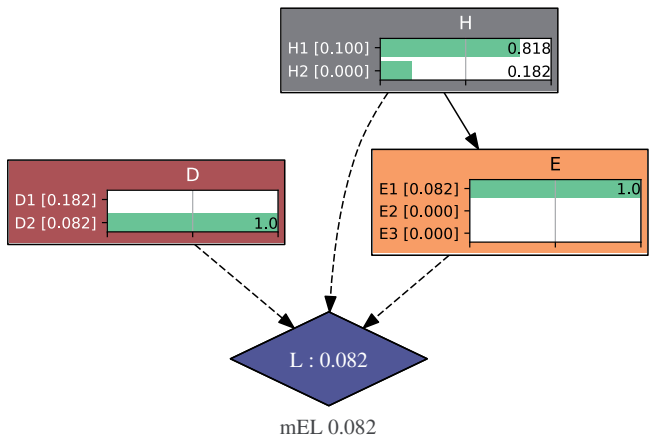

Figure 4. Illustration of the influence diagram shown in Fig. 1a, constructed in Python using aGrUM/pyAgrum (version 0.20.2). Node definitions are as given in Table 1, and quantitative assessments are as given in Section 2.4.4. The diagram illustrates a case where there is an output of type E_1 and the optimal decision is D_2 because its expected loss is smaller than that of decision D_1.

2.4.4 *Illustration by means of an influence diagram*

Consider the results presented in the previous sections in the form of an influence diagram, as shown in Fig. 4. This influence diagram has been created in Python using aGrUM/pyAgrum (version 0.20.2).[41] The model has the structure shown in Fig. 1a, although the positions of nodes E and H—as automatically arranged by the program—are slightly different. The influence diagram here shows a situation where expert output of type E_1 is available. This is conveyed by setting the state of node E to E_1. The node H shows posterior probabilities $\Pr(H_1|E_1)$ and $\Pr(H_2|E_1)$ of 0.818 and 0.182, respectively, obtained by updating a uniform prior[42] with a

[41] https://agrum.gitlab.io, see also Ducamp et al. (2020) for a description of this computing environment.
[42] That is $\Pr(H_1) = \Pr(H_2) = 0.5$.

likelihood ratio m/p of $0.9/0.2 = 4.5$. The probabilities $\Pr(H_j|E_i)$ are one ingredient in the computation of the expected loss as defined in Equation (1). Another is the loss function. For the purposes of illustration, we assume $\ell_1 = (E_1, H_2, D_1) = 1$ and $\ell_2 = (E_1, H_1, D_2) = 0.1$, that is a ratio of losses from adverse outcomes of 10. The remaining (optimal) outcomes are assigned a loss of zero. We can now obtain the expected loss for decision D_1, given E_1, using Equation (1), which reduces to $\mathrm{EL}(D_1|E_1) = \ell_1 \times \Pr(H_2|E_1)$. Following the assignments above, this gives $1 \times 0.182 = 0.182$ and is displayed next to the state D_1 of node D. Similarly, we obtain the expected loss for decision D_2, given E_1, as $\ell_2 \times \Pr(H_1|E_1)$. The result is $0.1 \times 0.818 = 0.0818$ and is displayed next to state D_2 of node D. Thus, in our example here, $\mathrm{EL}(D_2|E_1) < \mathrm{EL}(D_1|E_1)$, so D_2 is the better decision in terms of expected loss.

To put this result in context, consider Fig. 2b. This figure shows that for a loss ratio ℓ_1/ℓ_2 of 10, a posterior probability $\Pr(H_1|E_1)$ greater than 0.91 is required for decision D_1 to be the optimal decision. In other words, the posterior odds $\Pr(H_1|E_1)/\Pr(H_2|E_1)$ obtained here do *not* "outweigh" the loss ratio in the sense defined by Expression (3). As another way of looking at this result, consider Fig. 2a. This figure shows that for a posterior probability $\Pr(H_1|E_1)$ of 0.818, the y-axis value of the EL function of D_2 (dashed line) is *smaller* than that of D_1.

2.5 Modularity and flexibility

The modeling framework invoked in this article is both modular and flexible. It is modular in the sense that different aspects, in particular inference and decision, are captured by different but related sub-models. The framework is flexible in the sense that each sub-model can be refined as needed. To illustrate this feature, consider again the aspect of inference. In our model, Fig. 1a, we have chosen a minimal representation that includes only the nodes E (for expert output) and H (for the main propositions of interest). Other accounts in the literature have further dissected this relationship. For example, when dealing with human witnesses, a cascaded chain of inference can be constructed, relating the variable representing a witness's assertion about the occurrence of a given event to the variable representing that event through a series of intermediate nodes. These intermediate nodes represent propositions such as the witness's senses registering the target event and the witness actually believing that the event occurred (e.g. Taroni et al. 2014). Similarly, in the context of forensic DNA evidence, Thompson et al. (2003) distinguish between an expert's *report* of the occurrence of a correspondence between two compared DNA profiles, on the one hand, and the actual (but unobserved) event of corresponding DNA profiles, on the other. The latter includes considerations of the rarity of the features in the relevant population, referred to as diagnosticity (Koehler 2008), whereas the former accounts for the probability of (human) error, that is reliability (Koehler 2008). See, for example, Taroni et al. (2004) for a representation of this analytical view using a Bayesian network. It could be used in the models shown in Fig. 1 to replace the $H \rightarrow E$ fragment, thus illustrating the notions of modularity and flexibility. For further accounts of cascaded inference based on human sources of information, see also Schum (1994) and Thompson (2016).

3. Discussion and conclusions

The special paper section in this journal asks the question of what a future with practically implemented statistics-driven evaluation methods for forensic results, especially in fingerprint examination, might look like. We argue that this future will involve, more or less explicitly, some form of reliance. In this article, we have chosen to examine the notion of reliance through the lens of formal analysis. We derive a number of insights, which we summarize below.

Overall, our decision-analytic account of reliance on the output of an information source *in the instant case* differs from mainstream accounts of *general* performance assessment based on response rates in two-category classification problems, as largely described in the AI (machine learning) literature (e.g. Murphy 2012; Shalev-Shwartz and Ben-David 2014; Russell and Norvig 2016) and similarly adopted by black box studies in forensic comparison disciplines. Aggregate performance metrics, such as those derived from a confusion matrix, can be useful in method development, evaluation, and comparison, and can inform general discourses about the admissibility of particular methods. However, these metrics cannot be used directly to capture

specific expert output in a particular case. They have also been shown not to be conducive to understanding (Burnell et al. 2023). Our account clarifies why exactly aggregate performance metrics fall short of the needs in case-specific decision-making about reliance.

First, general performance metrics derived from test cases under controlled conditions are, by design, entirely data-based, making them seemingly objective and human independent. However, this falls short of the needs of decision-making in the individual case. What we mean by this is that reliance decisions, by definition, imply consequences of varying degrees of (un-)desirability, especially adverse consequences. In our context here, adverse consequences are reliance on inaccurate system output and nonreliance on accurate system output. What is particularly intricate is that the maker of the decision of reliance is *not* necessarily the party who has to bear the consequences of the decision, a consideration that is completely absent in controlled studies of forensic comparison disciplines such as fingermarks. Recognizing and quantifying the relative undesirability of decisional outcomes thus becomes a key feature of discourses over case-based decisions about reliance. This highlights the need to introduce value judgments, but data-centric performance metrics provide no help with this. The deeper insight here is that what characterizes the "goodness" of a method *in general* cannot serve as a criterion for characterizing the "goodness" of individual decisions (Biedermann et al. 2018). Thus, with respect to the future of machine-supported forensic feature comparison practice, our analysis suggests that (test) data *alone*, while useful for some purposes, is inherently insufficient to solve the problem of reliance in the instant case. Instead, as discussed below, there is an inherent judgmental component that ultimately relies on human input, even if that input was somehow programmed into a computational procedure.

Second, in our model, the expected loss associated with the decision of reliance in the individual case depends crucially on the probability of—in our context here—the accuracy of the expert output. It is important to note that we are talking about the accuracy of the individual output, not the accuracy of the method in general. Accuracy in a particular case necessarily depends on what evidence is available other than the expert output at hand, that is evidence that informs the decision-maker's beliefs about the relevant ground truth state. While the introduction of probability into decision-making procedures might be perceived as retrograde and to be avoided, we have shown its necessity. In particular, if we try to design a decision process based on avoiding excessive losses using a deterministic approach, this leads to paralysis. For example, in the case of system output that asserts support for a particular proposition rather than the relevant alternative, the use of an asymmetric loss function, reflecting a reasonable attitude, leads to the recommendation that one should *never* decide to rely on a particular type of expert output. Thus, if we want to allow for reliance decisions in the latter type of case, we cannot do so without accepting a nonzero probability of incurring the overall worst decision consequence. Our account makes this understanding formally precise. Qualitatively, the decision-theoretic criterion we have derived states that the higher the ratio of losses associated with the two ways in which reliance and nonreliance can lead to adverse outcomes, the more we shall be sure that the target ground truth state asserted by the system output is true, rather than the relevant alternative.

Ultimately, our account is agnostic about the type of expert output. It makes no difference whether the information comes from a traditional human expert, a machine, or a combination of the two. For this reason, our account alleviates common concerns about the potential epistemological challenges raised by the use of AI output, of which forensic fingerprint expertise is just one example. In terms of the logic of reliance decisions, our account emphasizes that there are essentially two components that need to be coherently aggregated: an expression of uncertainty about the relevant conditioning ground truth states, and an expression or judgment about the value (or, undesirability) of decision consequences. Neither of these can reasonably be externalized because, as noted above, they require a personal stance on the part of the decision-maker. We argue that these considerations should be at the heart of current and future discourses about the use of machine-generated or -assisted evidence, particularly fingerprint examination. However, our analysis in this article shows that even the seemingly simple question of reliance on information has no easy answers *if* a rationally rigorous procedure is to be devised.

The above result also provides a perspective on the broader question of whether decision-making can (justifiably) be delegated to AI systems. Critical readers may argue that this is a

nonissue, as there are already many operating systems that are entrusted with autonomous decision-making, such as one-to-one comparisons in biometric verification tasks. However, these applications typically operate with high-quality input information and are based on decision thresholds defined by performance metrics in the aggregate case, which are considered acceptable by the designers of such systems and by those who implement them. This is different from applications that make "one-to-many" comparisons, with possibly sub-optimal input information (i.e. a fingermark of poor quality), and that involve case-based value judgments and individualized assessments of uncertainty. In addition, the former applications (of type "one-to-one") implicitly have a built-in value judgment that the errors have tolerable consequences. As our article helps to illustrate, value judgments in one-to-many comparisons, possibly leading to associations between a POI (defendant) and a crime scene fingermark, obviously reflect completely different stakes (Kotsoglou and Biedermann 2022). In this sense, genuine delegation of decision-making tasks would require either the encoding of these core features of decision-making in the application, or the acceptance of a procedure based on a proxy for these features.

Acknowledgements

The authors thank Pierre-Henri Wuillemin from Sorbonne University, Paris, for his precious advice on the `aGrUM/pyAgrum` framework. The authors are indebted to the Guest Editors at *Law, Probability, and Risk*, the anonymous reviewers and Clare Lau of the Johns Hopkins University Applied Physics Laboratory for their valuable comments on the manuscript.

Funding

Alex Biedermann gratefully acknowledges the support of the Swiss National Science Foundation (grant no. BSSGI0_155809), the Société Académique Vaudoise, and the Fondation pour l'Université de Lausanne.

References

Aitken, C. G. G., Taroni, F. and Bozza, S. (2020) *Statistics and the Evaluation of Evidence for Forensic Scientists*, 3rd edn. Chichester: John Wiley & Sons.

Allen, R. J., and Miller, J. S. (1993) 'The Common Law Theory of Experts: Deference or Education?' *Northwestern University Law Review*, **87**: 1131–47.

Baron, J. (2008) *Thinking and Deciding*, 4th edn. New York: Cambridge University Press.

Baron, J. (2012) 'The Point of Normative Models in Judgment and Decision Making', *Frontiers in Psychology*, **3**: Article 577, 1–3.

Biedermann, A., Bozza, S., and Taroni, F. (2016) "The Decisionalization of Individualization', *Forensic Science International*, **266**: 29–38.

Biedermann, A., Bozza, S. and Taroni, F. (2020) 'Normative Decision Analysis in Forensic Science', *Artificial Intelligence and Law*, **28**: 7–25.

Biedermann, A., Bozza, S., Taroni, F., and Aitken, C. (2017) 'The Consequences of Understanding Expert Probability Reporting as a Decision', *Science & Justice, Special Issue on Measuring and Reporting the Precision of Forensic Likelihood Ratios*, **57**: 80–85.

Biedermann, A., Bozza, S., Taroni, F., and Garbolino, P. (2018) 'A Formal Approach to Qualifying and Quantifying the 'Goodness' of Forensic Identification Decisions', *Law, Probability and Risk*, **17**: 295–310.

Biedermann, A., Bozza, S., Taroni, F., and Vuille, J. (2020) 'Computational Normative Decision Support Structures of Forensic Interpretation in the Legal Process', *SCRIPTed: A Journal of Law, Technology and Society*, **17**: 83–124.

Biedermann, A., Caruso, D., and Kotsoglou, K. (2020) 'Decision Theory, Relative Plausibility and the Criminal Standard of Proof', *Criminal Law and Philosophy*, **15**: 131–57.

Biedermann, A., Garbolino, P., and Taroni, F. (2013) 'The Subjectivist Interpretation of Probability and the Problem of Individualisation in Forensic Science', *Science & Justice*, **53**: 192–200.

Biedermann, A., and Kotsoglou, K. N. (2021) 'Forensic Science and the Principle of Excluded Middle: "Inconclusive" Decisions and the Structure of Error Rate Studies', *Forensic Science International: Synergy*, **3**: 100147.

Biedermann, A., and Vuille, J. (2018a) 'The Decisional Nature of Probability and Plausibility Assessments in Juridical Evidence and Proof'. *International Commentary on Evidence*, **16**: 1–30.
Biedermann, A., and Vuille, J. (2018b) 'Understanding the Logic of Forensic Identification Decisions (Without Numbers)', *sui-generis*, **5**: 397–413.
Blackstone, W. (1769) *Commentaries on the Laws of England*, Vol. **4**. Oxford: Clarendon Press.
Bovens, L., and Hartmann S. (2003) *Bayesian Epistemology*. Oxford: Clarendon Press.
Brier, G. W. (1950) 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review*, **78**: 1–3.
Buckleton, J. S. et al. (2019) 'The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity', *Journal of Forensic Sciences*, **64**: 393–405.
Burnell, R. et al. (2023) 'Rethink Reporting of Evaluation Results in AI', *Science*, **380**: 136–8.
Castelvecchi, D. (2020) 'Beating Biometric Bias', *Nature*, **587**: 347–49.
Champod, C., Eldridge, H., and Lambert, S. (2020) *A Primer on Error Rates in Fingerprint Examination*. https://doi.org/10.5281/zenodo.3734560, accessed Apr. 2024.
Coble, M. D., and Bright, J.-A. (2019) 'Probabilistic Genotyping Software: An Overview', *Forensic Science International: Genetics*, **38**: 219–24.
Cole, S. A., and Biedermann, A. (2020) 'How Can a Forensic Result be a "Decision"? A Critical Analysis of Ongoing Reforms of Forensic Reporting Formats for Federal Examiners', *Houston Law Review*, **57**: 551–92.
Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.
de Finetti, B. (1974) *Theory of Probability, A Critical Introductory Treatment*, Vol. **1**. London: John Wiley & Sons.
Dekay, M. L. (1996) 'The Difference Between Blackstone-like Error Ratios and Probabilistic Standards of Proof', *Law & Social Inquiry*, **21**: 95–132.
Dror, I. and Mnookin, J. L. (2010) 'The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science', *Law, Probability and Risk*, **9**: 47–67.
Ducamp, G., Gonzales, C., and Wuillemin, P.-H. (2020) aGrUM/pyAgrum: a toolbox to build models and algorithms for probabilistic graphical models in Python. In *10th International Conference on Probabilistic Graphical Models*, Skørping, Denmark, pp. 609–612.
Faigman, D. L., Monahan, J., and Slobogin, C. (2014) 'Group to Individual (G2i) Inference in Scientific Testimony', *The University of Chicago Law Review*, **81**: 417–80.
Friedman, R. D. (2018) 'The Persistence of the Probabilistic Perspective', *Seton Hall Review*, **48**: 1589–600.
Garrett, B., and Monahan, J. (2019) 'Assessing Risk: The Use of Risk Assessment in Sentencing', *Judicature*, **103**: 42–9.
Garrett, B., and Monahan, J. (2020) 'Judging Risk', *California Law Review*, **108**: 439–93.
Gittelson, S., Bozza, S., Biedermann, A. and Taroni, F. (2013) 'Decision-theoretic Reflections on Processing a Fingermark', *Forensic Science International*, **226**: e42–7.
Good, I. J. (1950) *Probability and the Weighing of Evidence*. London: Griffin.
Hofmann, H., Carriquiry, A. and Vanderplas, S. (2020) 'Treatment of Inconclusives in the AFTE Range of Conclusions', *Law, Probability and Risk*, **19**: 317–64.
Imwinkelried, E. J. (2020) 'The Admissibility of Scientific Evidence: Exploring the Significance of the Distinction Between Foundational Validity and Validity as Applied', *Syracuse Law Review*, **70**: 817–49.
Jacquet, M., and Champod, C. (2020) 'Automated Face Recognition in Forensic Science: Review and Perspectives', *Forensic Science International*, **307**: 110124.
Kaplan, J. (1968) 'Decision Theory and the Factfinding Process', *Stanford Law Review*, **20**: 1065–92.
Kaye, D. H. (1987) 'The Validity of Tests: Caveant Omnes', *Jurimetrics Journal*, **27**: 349–61.
Kaye, D. H. (1999) 'Clarifying the Burden of Persuasion: What Bayesian Decision Rules do and do not do', *The International Journal of Evidence & Proof*, **3**: 1–29.
Kjærulff, U. B., and Madsen, A. L. (2008) *Bayesian Networks and Influence Diagrams, A Guide to Construction and Analysis*. New York: Springer.
Koehler, J. J. (2008) 'Fingerprint Error Rates and Proficiency Tests: What They are and Why They Matter', *Hastings Law Journal*, **59**: 1077–100.
Koehler, J. J. (2017) 'Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences', *Arizona State Law Journal*, **49**: 1369–416.
Kotsoglou, K. N., and Biedermann, A. (2022) 'Inroads into the Ultimate Issue Rule? Structural Elements of Communication Between Experts and Fact Finders', *The Journal of Criminal Law*, **86**: 223–40.
Lau, T., and Biedermann, A. (2020) 'Assessing AI Output in Legal Decision-making with Nearest Neighbors', *Penn State Law Review*, **124**: 609–55.
Lindley, D. V. (1985) *Making Decisions*, **2**nd edn. Chichester: John Wiley & Sons.

Lindley, D. V. (1987) 'The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems', *Statistical Science*, **2**: 17–24.
Lindley, D. V. (2006) *Understanding Uncertainty*. Hoboken: John Wiley & Sons.
Mejia, R., Cuellar, M., and Salyards, J. (2020) 'Implementing Blind Proficiency Testing in Forensic Laboratories: Motivation, Obstacles, and Recommendations', *Forensic Science International: Synergy*, **2**: 293–8.
Mnookin, J. L. (2008) 'Expert Evidence, Partisanship, and Epistemic Competence', *Brooklyn Law Review*, **73**: 1009–33.
Morrison, G. S. et al. (2021) 'Consensus on Validation of Forensic Voice Comparison', *Science & Justice*, **61**: 299–309.
Murphy, K. P. (2012) *Machine Learning: a Probabilistic Perspective*. Cambridge: MIT Press.
Nunn, A. (2019–2020) 'Machine-generated Evidence', *TheSciTechLawyer*, **16**: 4–7.
PCAST (2016) *President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, D.C.: Executive Office of the President.
Raiffa, H. (1968) *Decision Analysis, Introductory Lectures on Choices under Uncertainty*. Reading, Massachusetts: Addison-Wesley.
Ramos, D., and Gonzalez-Rodriguez, J. (2013) 'Reliable Support: Measuring Calibration of Likelihood Ratios', *Forensic Science International*, **230**: 156–69.
Roth, A. (2016) 'Trial by Machine', *The Georgetown Law Journal*, **104**: 1245–305.
Roth, A. (2017) 'Machine Testimony', *The Yale Law Journal*, **126**: 1972–2053.
Russell, S., and Norvig, P. (2016) *Artificial Intelligence. A Modern Approach*, 3rd edn. Essex: Pearson Education Ltd.
Schum, D. A. (1994) *Evidential Foundations of Probabilistic Reasoning*. New York: John Wiley & Sons, Inc.
Shalev-Shwartz, S. and Ben-David, S. eds. (2014) *Understanding Machine Learning, From Theory to Algorithms*. Cambridge: Cambridge University Press.
Shenoy, P. P. (1992) 'Valuation-based Systems for Bayesian Decision Analysis', *Operations Research*, **40**: 463–84.
Shinkins, B., Thompson, M., Mallett, S., and Perera, R. (2013) 'Diagnostic Accuracy Studies: How to Report and Analyse Inconclusive Test Results', *British Medical Journal*, **346**: f2778.
Swofford, H., and Champod, C. (2021) 'Implementation of Algorithms in Pattern & Impression Evidence: A Responsible and Practical Roadmap', *Forensic Science International: Synergy*, **3**: 100142.
Swofford, H. J., Cole, S. A., and King, V. (2021) 'Mt. Everest—We Are Going to Lose Many: A Survey of Fingerprint Examiners' Attitudes Towards Probabilistic Reporting', *Law, Probability and Risk*, **19**: 255–91.
Swofford, H. J. et al. (2018) 'A Method for the Statistical Interpretation of Friction Ridge Skin Impression Evidence: Method Development and Validation', *Forensic Science International*, **287**: 113–26.
Swofford, H. J. et al. (2024) 'Inconclusive Decisions and Error Rates in Forensic Science', *Forensic Science International: Synergy*, **8**: 100472.
Taroni, F. et al. (2014) Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science, 2nd edn. *Statistics in Practice*. Chichester: John Wiley & Sons.
Taroni, F., Biedermann, A., Garbolino, P., and Aitken, C. G. G. (2004) 'A General Approach to Bayesian Networks for the Interpretation of Evidence', *Forensic Science International*, **139**: 5–16.
Taroni, F., Bozza, S., Biedermann, A., and Aitken, C. (2016) 'Dismissal of the Illusion of Uncertainty in the Assessment of a Likelihood Ratio', *Law, Probability and Risk*, **15**: 1–16.
Tart, M. (2020) 'Opinion Evidence in Cell Site Analysis', *Science & Justice*, **60**: 363–74.
Thompson, W. C. (2016) 'Determining the Proper Evidentiary Basis for an Expert Opinion: What do Experts Need to Know and When do They Know too Much?' in C. Robertson and A. Kesselheim (eds), *Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science, and Law*, pp. 133–150. Amsterdam: Academic Press.
Thompson, W. C., Taroni, F., and Aitken, C. G. G. (2003) 'How the Probability of a False Positive Affects the Value of DNA Evidence', *Journal of Forensic Sciences*, **48**: 47–54.
U.S. Department of Justice (2020) *Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline*, vers. 8.15.20. https://www.justice.gov/olp/page/file/1284786/, accessed Apr. 2024.
von Winterfeldt, D., and Edwards, W. (1986) *Decision Analysis and Behavioral Research*. Cambridge: Cambridge University Press.
Zabell, S. (2012) 'Book Review: Statistical DNA Forensics: Theory, Methods and Computation, by Wing Kam Fung and Yue-Qing Hu', *Law, Probability and Risk*, **11**: 105–10.

Appendix 1: Deciding about reliance in case of expert output of type E_3

For the sake of completeness, we include a brief note on the properties of the decision criterion for the case of expert output E_3 (asserted support for H_2 over H_1). First, we give the expected loss for each of the two decisions D_1 and D_2, using Equation (1). For decision D_1, we obtain:

$$\mathrm{EL}(D_1|E_3) = \underbrace{\mathrm{L}(E_3, H_1, D_1)}_{\ell_3} \times \Pr(H_1|E_3) + \underbrace{\mathrm{L}(E_3, H_2, D_1)}_{0} \times \Pr(H_2|E_3) = \ell_3 \times \Pr(H_1|E_3)$$

The expected loss of decision D_2 is:

$$\mathrm{EL}(D_2|E_3) = \underbrace{\mathrm{L}(E_3, H_1, D_2)}_{0} \times \Pr(H_1|E_3) + \underbrace{\mathrm{L}(E_3, H_2, D_2)}_{\ell_4} \times \Pr(H_2|E_3) = \ell_4 \times \Pr(H_2|E_3)$$

To investigate the conditions under which D_1 is preferable to D_2, we need to examine the conditions under which $\mathrm{EL}(D_1|E_3) < \mathrm{EL}(D_2|E_3)$. Reformulating this expression using the above results and a slight rearrangement of the terms leads to a result similar to Expression (3):

$$\Pr(H_1|E_3)/\Pr(H_2|E_3) < \ell_4/\ell_3. \tag{A.1}$$

Note an important difference between this result, for output of type E_3, and Expression (3) for the case of expert output of type E_1. The result here is that for D_1 to have a smaller expected loss than D_2, the posterior odds of $\Pr(H_1|E_3)$ to $\Pr(H_2|E_3)$ must be *smaller* than the ratio of the losses associated with the adverse consequences of the decisions of nonreliance D_2 and reliance D_1, respectively. Figure A.1 (a) and (b) summarizes this result.

For example, suppose that ℓ_4, the loss of nonreliance on E_3 when H_2 is true (i.e. the output is accurate), is 10 times greater than ℓ_3, the loss associated with reliance on E_3 when H_1 is true (i.e. the output is not accurate): $\ell_4/\ell_3 = 10$. In such a situation, decision D_1 is preferable to decision D_2 in terms of expected loss *only as long as* the posterior odds $\Pr(H_1|E_3)/\Pr(H_2|E_3)$ is *smaller* than 10, that is the posterior probability $\Pr(H_1|E_3)$ is smaller than 0.91. This is an intricate result. It says that E_3 (asserted support for H_2 over H_1) should be relied upon *even if* the odds are in favor of H_1 over H_2 (i.e. the probability of E_3 being inaccurate is greater than 0.5). Nonreliance on E_3, decision D_2, becomes the better decision only when the odds exceed a certain level, that is "become very high" (informally speaking). This is a consequence of our preference structure (loss function): we consider it *less* undesirable to rely on falsely exculpatory

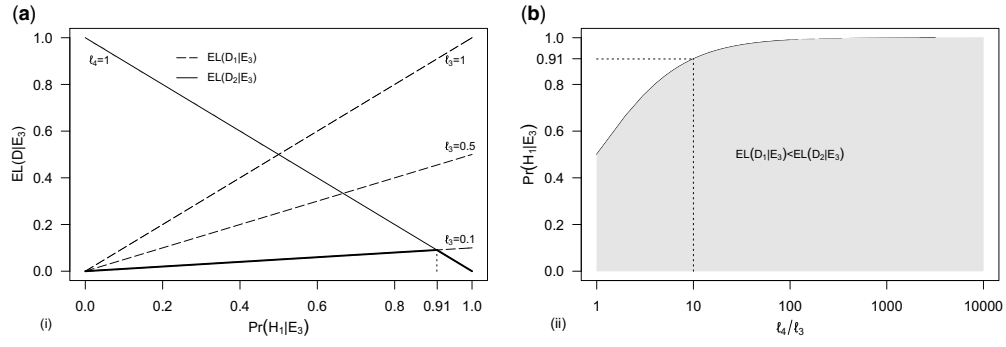

Figure A.1. (**a**) Expected loss of decisions D_1 (reliance on expert output E_3; dashed line) and D_2 (not relying; solid line) as a function of $\Pr(H_1|E_3)$, using a 0 – 1 loss function with $l_3 = \{0.1, 0.5, 1\}$ (loss of reliance on E_3 when H_1 is true) and $l_4 = 1$ (loss of nonreliance E_3 when H_2 is true). The bold line highlights the optimal decision as a function of $\Pr(H_1|E_3)$ for a case where ℓ_4 is 10 times greater than ℓ_3. (**b**) Maximum probability $\Pr(H_1|E_3)$ so that decision D_1 has the smaller expected loss than decision D_2, as a function of the loss ratio ℓ_4/ℓ_3. The gray-shaded area shows pairs of values of loss ratio and probability $\Pr(H_1|E_3)$ for which D_1 minimizes expected loss.

output (in the case of E_3, deciding D_1 when H_1 is the case) than not to rely on correctly exculpatory output (deciding D_2 when H_2 is the case for output E_3).

It may be tempting to conclude that the above result is banal, but this may only be because it corresponds well to our intuition. The less obvious point is to show, through formal analysis, that a particular result (or intuition) actually has defensible and logical grounds, which is the aim of our development here.

Law, Probability and Risk, 2024, 00, 1–28
https://doi.org/10.1093/lpr/mgae007
Research Article