Forensic Science International: Synergy 3 (2021) 100147

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

Forensic science and the principle of excluded middle: "Inconclusive" decisions and the structure of error rate studies

Alex Biedermann [a, *], Kyriakos N. Kotsoglou [b]

[a] University of Lausanne, School of Criminal Justice, 1015 Lausanne-Dorigny, Switzerland
[b] Northumbria University, School of Law, Newcastle Upon Tyne, NE1 8ST, United Kingdom

ARTICLE INFO

Article history:
Received 23 February 2021
Received in revised form
20 March 2021
Accepted 29 March 2021
Available online 17 April 2021

Keywords:
Inconclusive
Decision
Error rate
Principle of excluded middle

ABSTRACT

In a paper published recently in this journal, Dror and Scurich (2020) [20] critically discuss the notions of "inconclusive evidence" (i.e., test items for which it is difficult to render a categorical response) and "inconclusive decisions" (i.e., experts' conclusions or responses) in the context of forensic science error rate studies. They expose several ways in which the understanding and use of "inconclusives" in current forensic science research and practice can adversely affect the outcomes of error rate studies. A main cause of distortion, according to Dror and Scurich, is what they call "erroneous inconclusive" decisions, in particular the lack of acknowledgment of this type of erroneous conclusion in the computation of error rates. To overcome this complication, Dror and Scurich call for a more explicit monitoring of "inconclusives" using a modified error rate study design. Whilst we agree with several well-argued points raised by the authors, we disagree with their framing of "inconclusive decisions" as potential errors. In this paper, we argue that referring to an "inconclusive decision" as an error is a contradiction in terms, runs counter to an analysis based on decision logic and, hence, is questionable as a concept. We also reiterate that the very term "inconclusive decision" disregards the procedural architecture of the criminal justice system across modern jurisdictions, especially the fact that forensic experts have no decisional rights in the criminal process. These positions do not ignore the possibility that "inconclusives" – if used excessively – do raise problems in forensic expert reporting, in particular limited assertiveness (or, overcautiousness). However, these drawbacks derive from inherent limitations of experts rather than from the seemingly erroneous nature of "inconclusives" that needs to be fixed. More fundamentally, we argue that attempts to score "inconclusives" as errors amount to philosophical claims disguised as forensic methodology. Specifically, these attempts interfere with the metaphysical substrate underpinning empirical research. We point this out on the basis of the law of the excluded middle, i.e. the principle of "no third possibility being given" (*tertium non datur*).

"The statement of Heraclitus, that everything is and is not, seems to make everything true, but that of Anaxagoras, that an intermediate exists between two contradictories, makes everything false; for when things are blended, the blend is neither good nor not-good, so that it is not possible to say anything truly."
Aristotle [2, Book IV, 1012a25-29.]

1. Introduction

A key concern for legal officials who deal with expert witness testimony is whether these witnesses can deliver on their basic promise, i.e. to act as (proper) experts in the respective empirical field. This raises the immediate question on what to tie an expert's status. Pragmatic fact-finders may find resumes a convenient source of information, and useful as a minimum check, yet this comes with no guarantees. In the United States, the response to this problem has for a long time been the "general acceptance" test set out in *Frye*,[1] superseded (as regards the federal courts and some

* Corresponding author. University of Zürich, Faculty of Law, 8006, Zürich, Switzerland.
E-mail address: alex.biedermann@unil.ch (A. Biedermann).

[1] *Frye v. United States* 293 F. 1013 (D.C.Cir. 1923).

https://doi.org/10.1016/j.fsisyn.2021.100147

state courts) by *Daubert*,[2] which engineered a responsibility shift away from the scientific community towards trial judges who are now seen as gatekeepers of scientific validity. Interestingly, these cases continue to echo widely even in jurisdictions in which they have no direct applicability. Besides jurisprudence, the fundamental requirement of qualification and proficiency, or at least specialised knowledge, pervades legislation,[3] scholarly writing [e.g., 13, 38] and reports by eminent commissions [e.g., 35].

Part of *Daubert*'s non-definitive toolkit is, among other aspects, the "known or potential rate of error".[4] This notion of error rates has been discussed widely and with respect to entire fields of forensic practice, such as friction ridge mark (i.e., "fingerprint") examination [27]. However, since any forensic field necessarily involves the activities of human experts, the notion of error (rate) cannot be dissociated from empirical performance measures on the level of individual examiners as revealed, for example, through tests conducted under controlled conditions. In these contexts, various types of studies are commonly mentioned, in particular accuracy or validation studies, but also proficiency tests, though the latter are less amenable to generating performance characteristics that generalize to a method or technique as a whole. Broadly speaking, the idea here is an empirical warrant for experts' claims of proficiency, most typically in the form of a metric, such as error rates [e.g., 27, 32]. The PCAST report highlights this understanding by emphasising the importance of so-called blackbox studies. These are studies "in which many examiners render decisions about many independent tests (typically, involving 'questioned' samples and one or more 'known' samples) and the error rates are determined" [35, at p.5–6].

Proficiency scores of individual examiners can, in principle, be obtained in the same way, by having examiners keep records of their scores and disclosing documentation of such data upon request [11]. For example, in a recent letter to Judge Patrick Schiltz, Chair of the Advisory Committee on the Federal Rules of Evidence, the U.S. Department of Justice noted:

> "The Department recognizes that a forensic examiner's past performance on relevant, skill-based testing is an important measure for evaluating her performance in a given case. As such, FBI proficiency test results are routinely provided to defense counsel upon request. The FBI Laboratory will soon begin disclosing proficiency test results without a specific defense request as part of their general discovery and disclosure procedures." [17, at p.6]

At first glance, error rates seem a straightforward concept. For example, when evaluating the performance characteristics of a diagnostic test for a medical condition, a study is conduced on truly diseased and truly nondiseased individuals. Similarly, in forensic science, one can consider pairs of test items for which it is known whether they come from the same source or not. By asking examiners to assert, for each pair of test items, whether they come from the same source or from different sources, one *could* readily keep track of the number of accurate and erroneous responses across all examiners, but also for individual examiners. However, the difficulties lie in the details. The analogy to the method of establishing performance metrics in other fields, in particular medical applications, provides some insightful analogies.

Consider the example of a medical diagnostic test for detecting a target substance, such as sperm [1]. The test items are designed in a way such that the target substance is either present in the examined item (verifiable by a reference method), or absent. These are the two mutually exclusive conditions. The test outcome is dichotomous, it turns out to be either positive or negative, though this is a simplification to which we come back in due course. How good a test is in detecting the presence or absence of the target substance in test items can routinely be characterised by the following standard performance measures. The proportion of positive test results among the items that truly contain the target substance provides the true positive rate (or, sensitivity). The proportion of positive test results among the items that do not contain the target substance informs us about the false positive rate. In turn, the proportion of negative test results among the items that do not contain the target substance provides the specificity of the test.[5] Note, however, that this is an idealistic view. In practice, diagnostic test outcomes may not be univocally categorised as either positive or negative [29], for various reasons (see, e.g. Ref. [39], for a review).

Now compare this setup to a typical forensic identification problem. In a forensic identification setting, there are also two mutually exclusive and exhaustive propositions: either the examined trace (e.g., a fingermark) comes from the person of interest, or the trace comes from an unknown person. These two possibilities are commonly referred to as same- and different source propositions (or, specific versus unknown source propositions). Thus far, the problem is in line with the medical test example [15, at p.568]. The analogy also holds when it comes to the result (test outcome), i.e. examiners' conclusions (i.e., responses): forensic scientists do not necessarily state an identification or an exclusion. Many forensic scientists operate on a tripartite framework including the response "inconclusive" [e.g., 18].

This raises the question of how to properly score this type of conclusion so as not to distort error rates. The purpose of this paper is to analyse and discuss opposing views regarding this question. Relevant for our discussion is a recent paper by Dror and Scurich [20] because it critically exposes problems in the way "inconclusive" decisions tend to be dealt with in current forensic science research and practice. Dror and Scurich [20] rightly note, for instance, that there is potential for examiners to misuse "inconclusive" as a response category to artificially improve their performance (as measured by the error rate). Specifically, by conveniently avoiding to count "inconclusives" as errors, an examiner can cook the books by resorting to this category every time a definite 'call' (of either identification or exclusion) cannot easily be made, and yet incur no penalty in subsequent scoring. To improve the current study designs for performance assessment, Dror and Scurich [20] recommend a method for scoring "inconclusives", previously proposed in Ref. [16].[6] Under certain conditions, this method treats "inconclusives" as errors. While it is difficult to disagree with Dror and Scurich [20] that current practices for processing "inconclusives" are unsatisfactory, and prone to adversely affect standard procedures for computing error rates, the proposed remedies only compound and shift the problem. We will critically review standard terminology used in this context to illustrate this point. Methodologically, our analysis draws on decision logic and evidence law doctrine, mainly because forensic examiners' conclusions are now widely referred to as "decisions" [15].

[2] *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 U.S. 579 (1993).

[3] See e.g. FRE 702 (USA) which states that a witness "is qualified as an expert by knowledge, skill, experience, training, or education". Similarly, the Criminal Procedure Rule 19.4 (E + W) requires that an expert's report must "give details of the expert's qualifications, relevant experience and accreditation".

[4] *Daubert*, supra note 2, at 594.

[5] See also [26] for a detailed discussion of performance metrics for medical and psychological tests with a particular focus on applications in legal contexts.

[6] Note, however, that in Ref. [16] the method was considered for a slightly different purpose, i.e. casework only.

This paper is structured as follows. Section 2 starts with an overview of common understandings of the notion of "inconclusive" as used by forensic scientists and the problem these understandings cause in error rate studies. This section also examines arguments for and against treating "inconclusives" as potential errors and exposes drawbacks of the modified error rate study design proposed in Ref. [20]. Section 3 reviews the controversy over "inconclusives" from a decision-theoretic point of view. We explain the statements regarding the notion of error that, in a decision-theoretic account, are and are not warranted. We also include a doctrinal analysis that exposes the incongruence between the understanding of expert conclusions as decisions and the notion of decisional rights in a legal perspective. Conclusions are presented in Section 4.

2. "Inconclusive" decisions in mainstream error rate doctrine: problems and suggested solutions

2.1. Preliminaries

To start with, it is important to note that the response category "inconclusive" is part of a discrete, tripartite reporting vocabulary, alongside the conclusions "(source) identification" and "(source) exclusion" [e.g., 18]. Note, however, that regarding "inconclusives" as a topic worthy of discussion does not mean that we recognise or approve this reporting scheme as a suitable framework. As we will explain in the forthcoming sections, the reasons why we do not subscribe to this reporting framework are both methodological and doctrinal in nature. From a legal point of view, the procedural architecture of the criminal justice system invalidates any decisions made by forensic experts (see Section 3.2). The arguments pertaining to forensic science are well documented in literature [see, e.g., 10, 14, 41] and need not be reiterated here. It suffices to note that there now are logical methods for evaluating scientific evidence and expert testimony [22]. These evaluation methods abstain from opining directly on competing propositions (i.e., same and different source), and come with concepts to measure the performance of their output [37]. Rather than on propositions, these methods focus on the probative value of the *evidence* (e.g., in terms of a likelihood ratio). The notion of "inconclusive" is also sometimes used in this context, though only to refer to evidence that is neutral (i.e. not probative in one way or another) and without raising problems in the method's performance assessment. For the time being, however, reporting language involving the term "inconclusive" – in the sense of an opinion on a proposition – is still widely used by forensic practitioners across different forensic domains and across jurisdictions, despite legal prohibitions and methodological flaws. Thus, problems related to the use of "inconclusives" and drawbacks of proposed remedies for those problems rightfully merit attention.

2.2. The meaning of concluding "inconclusive"

The DOJ's ULTR defines "inconclusives" as follows [18, at p.3; emphasis added]:

> 'Inconclusive' is an examiner's *conclusion* that there is insufficient quantity and/or clarity of corresponding friction ridge skin features between two impressions such that the examiner is unable to identify or exclude the two impressions as originating from the same source.
> The basis for an 'inconclusive' conclusion is an examiner's opinion that a 'source identification' or 'source exclusion' cannot be made due to insufficient information in either of the two impressions examined.

Note the term 'opinion' in the second paragraph, which is a novelty in the current version of the DOJ's ULTR,[7] as compared to the previous version that used the term 'decision' [15]. Whether treating an "inconclusive" as an opinion is more suitable than treating it as a decision is a discussion of its own. In the next parts of this paper, we will highlight some insights that the term "decision" can provide, as compared to "opinion", and explain why we maintain that the term decision suitably captures the nature of "inconclusives" without running into contradictions [8,9]. The basic message conveyed by the DOJ's ULTR, however, is clear: when reporting "inconclusive", one does neither identify nor exclude [27]. While the focus of this type of statement is on ground truth, which forensic examiners should avoid [45], the notion of "inconclusive" can also be used with an emphasis on the evidence (see also Section 2.1): i.e., the statement that the observations (or, findings) do not help to discriminate between competing propositions regarding ground truth (i.e., same or different source). For example, OSAC's proposed Standard for Friction Ridge Examination Conclusions defines "inconclusive" as "the conclusion that the observations do not provide a sufficient degree of support for one proposition over the other" [34].

These considerations refer to the *information* that is conveyed when examiners use the term "inconclusive" in their reporting. It is fair to say that this interpretation is broadly accepted throughout forensic science. In turn, a separate and contentious issue is how "inconclusives" should be *treated*, especially whether they should be considered as errors [16]. In some sense, an "inconclusive" could be regarded as an error because the examiner fails to assert whether or not the compared items come from the same source. Yet, at the same time, by not asserting one of the two ground truth states, no misleading and hence no erroneous assertion is made. These two viewpoints thus refer, in opposing ways, to the congruence between the "inconclusive" conclusion issued by the examiner and the actual ground truth.

Some discussants insist that the erroneous nature of "inconclusive" decisions is revealed by the consequences that these decisions entail. In particular, so the argument, an "inconclusive" decision may "neutralize" [16, at p.816] vital evidence and, thus, deprive prosecution and defense from inculpatory and exculpatory evidence, respectively. Conversely, "inconclusives" may also be viewed in a more favourable light as erring on the safe side, or as an expression of caution. Specifically, allowing examiners to make a "pass" [27] provides them with a way to manage the potential for false identifications and false exclusions. However, the crux of the matter is the definition of the extent to which resorting to "inconclusive" is acceptable, i.e. not leading to distortions.

Thus, while there are robust *general* arguments for and against considering "inconclusives" as potential errors, there is a highly sensitive area where their differential treatment has consequences that represent a serious cause of concern. This area includes error rate studies, a field in which the processing of "inconclusives" has critically been discussed in a recent paper by Dror and Scurich [20]. We now turn to this paper.

2.3. How "inconclusives" can distort error rates

Studies on error rates commonly set aside "inconclusives" by not counting them as errors on grounds that there is no agreeable way of asserting when an "inconclusive" is to be considered an appropriate conclusion [e.g., 21]. Critical commentators regard this

[7] We are grateful to Professor Simon Cole for drawing our attention to this change in terminology.

position as problematic because, so their argument, this approach has the potential to distort error rates. Specifically, Dror and Scurich note [20, at p.334–335]:

> If one refuses a priori to count inconclusive decisions as errors, then error rates may be artificially and falsely reduced by making inconclusive decisions. In fact, zero error rates are possible with such an approach: regardless of anything, just reach inconclusive decisions for every comparison and you will have a perfect score!

Strictly speaking, Dror and Scurich are right on their point, but their critique depicts an extreme extrapolation of an abusive reporting policy that is unlikely to be endorsed by, let alone be useful to, examiners. Suppose that an examiner would in fact provide, as Dror and Scurich suggest, only "inconclusives" when being tested under controlled conditions of the conventional study design, and thus be credited by a perfect score. This achievement would come at a high price: the examiner would make him/her known as someone who *never* provides an identification or exclusion. Stated otherwise, such an examiner would be perfectly unhelpful and run out of business. Possibly, one might prefer calling such an expert in order to neutralize evidence that is probative in one way or another, but this seems far-fetched. We agree, however, that their example intends to illustrate the bias in the strategy of overusing "inconclusive", as any less than extreme use of "inconclusive" also tends to artificially decrease the error rate. An additional, realistic concern is that examiners who excessively employ "inconclusives" in proficiency testing, to embellish their performance, may not exhibit the same propensity towards reporting "inconclusives" in casework, thus adding a further dimension in which error rates can be uninformative [20].

Technically speaking, we share the above concerns about the potential for misuse, yet a question of interest is whether the real source of the these problems is the convention of never counting "inconclusives" as potential errors. Where "inconclusives" are excessively used and imply "negative consequences, such as misrepresenting error rate estimates in court which are artificially low and inaccurate" [20, at p.335], it is legitimate to ask whether this observation is actually the fault of the concept of "inconclusives", or whether it is a problem of corrupt (forensic) science in the first place. This is a rather speculative question and its answer might be immaterial to parties who see their evidence unduly neutralised by an "inconclusive" decision. Yet, from a methodological point of view, clarity on foundational principles is paramount for devising coherent practical proceedings. The questions of whether and how to treat "inconclusives" as potential errors are examples of questions that ask for such clarification. Answering them is important for moving forward mainly because counting "inconclusives" as potential errors is not without complications.

2.4. Recurrent arguments for and against treating "inconclusives" as potential errors

As outlined in Section 2.2, it is relatively easy to find prima facie plausible stances for and against considering "inconclusives" as potential errors. On the one hand, since the response "inconclusive" explicitly abstains from asserting one of the two possible ground truths (i.e., same or different source), it cannot, by definition, be an error. On the other hand, proponents of treating "inconclusives" as potential errors argue that it is a failure *not* to make a categorical assertion regarding ground truth where such an assertion could have been made. For example, Dror and Scurich note: "It depends on whether or not there is sufficient quality and quantity of information to reach an identification or exclusion decision" [20, at p.335]. Invoking this, however, culminates in changing the definition of the reference point. That is, Dror and Scurich widen the definition of error to the extent that they are able to make a case for scoring "inconclusives" as errors. They write: "the question is not only about the ground truth of who left the mark, but more about what is the correct conclusion given the information available in the evidence" [20, at p.334]. Note that this differs from the conventional definition of error that relies *only* on the congruence between an examiner's assertion and actual ground truth (see also Section 1). However, Dror and Scurich provide no independent argument for inflating the definition of error in the first place. Likewise, it is not clear why the question of interest should actually be "*more* about what is the correct conclusion given the information available in the evidence" [20, at p.334; emphasis added], rather than about ground truth.

One might object to this by arguing that the need to extend the definition of error stems from the fact that "inconclusives" can be used to embellish error rates. However, this places the cart before the horse. As we have noted in Section 2.3, the mere fact that "inconclusives" can be misused does not imply that the fault lies on the side of the concept of "inconclusives" and their scoring. It is important to avoid misconstruing a *perceived* inconvenience in the practical deployment of "inconclusives" as a definitional deficiency. Likewise, no one would condemn or even give up on arithmetic in finance simply because sometimes numbers are manipulated to commit fraud.

Dror and Scurich invoke the practical observation that different examiners might give different conclusions regarding the same comparison, and that sometimes even the same examiner gives different conclusions regarding the same comparison at different instances of time – and that this is not only unsatisfactory, but that this also implies that some examiners err. They note: "If some examiners conclude an identification (or exclusion) whereas other examiners conclude an inconclusive, then at least some of the examiners are mistaken" [20, at p.334] and "it cannot be that *all* examiners are correct when they reach *different* conclusions on the same evidence" [19, at p.704; emphasis as in original]. To be clear, if one takes ground truth as the reference point (here: same or different source), then categorical expert assertions of identification and exclusion that do not correspond to ground truth, are errors. In this sense only either identification or exclusion is a correct answer. However, concluding from this that examiners who report "inconclusive" while others identify or exclude are in error, is not straightforward. What the empirical observation of some examiners reporting "inconclusive" means is that, first of all, some examiners are less *assertive* than others. But scoring the less assertive expert as erroneous would require one to assume that there can only be one correct answer. This is an assumption for which Dror and Scurich provide no independent justification. Instead, their position is axiomatic in that it refers back to their extended definition of error given above, predicated on the idea of there being a (single) "correct conclusion given the information available in the evidence" [20, at p.334].

Let us pause here for a moment to clarify what this shifting definition of error implies. An immediate consequence is that the criterion for determining whether a given conclusion is an error is no longer based on ground truth (which can be easily controlled for in experimental settings). Instead, the determination will be based on an explicitly ground truth-independent status of what *ought* to be "the correct conclusion given the information available in the evidence" [20, at p.334]. This implies nothing less than defining category labels based on human judgment (either as a consensus or majority vote [20]) regarding perceivable features of the evidence ("quality and quantity of information" [20, at p.335]) *other than ground truth*. Let us recall that ground truth in the traditional sense

refers to an actual state of nature and a fixed truth-value for propositions salient in a given experiment: e.g. the compared items either come or do not come from the same source. Defining errors with respect to such real states of nature enshrines the notion of factual accuracy which, by the way, pervades the legal system, including determinations such as legal verdicts. Dror and Scurich's view, however, conflates the ontological level of analysis (where ground truth is fixed) with the epistemic level of analysis (where ground truth remains uncertain). Thus, they conflate the *performance characteristics* of a test or method (i.e., sensitivity and specificity) with the *epistemic features* of expert witness testimony. This is equivalent to arguing for a third, intermediate stage of pregnancy, treating this condition as a non-binary, as a result of an imperfect pregnancy test. But, we cannot change the meaning of words and their possible ontological states due to inherently imperfect measuring instruments, and we cannot allow uncertainty to escape the epistemic domain and infiltrate the ontological domain. As an aside, it is interesting to note that diagnostic accuracy studies in medical literature also consider only two ground-truth states (the so-called "target condition" [12]) and three response categories, leading to the well-known 3×2 tables [23,40].[8]

For forensic purposes, ground truth in controlled experiments (e.g., validation studies or proficiency tests) can, in principle, unequivocally be established.[9] In turn, Dror and Scurich suggest that this reference point should be replaced by category labels established exclusively by some sort of inherently unequivocal forensic wisdom that takes the form of either a *Frye*esque-consensus among independent experts, or a majority vote. This manages to miss the basic lesson from *Daubert*: consensus in the respective community is simply a surface feature of established and robust protocols and methods, not their core feature. Methods are not sound *when* or *because* experts agree on them. On the contrary, there is scientific consensus when these methods exhibit particular levels of performance. Arguing otherwise confuses cause and effect by reducing scientific status and reliability to consensus or decision-making rules (e.g. majority vote) rather than to methodological features.

In practice, Dror and Scurich's framework tends to encourage individual experts to no longer think in terms of whether their conclusion in the instant case aligns with actual ground truth. Instead, they are directed towards thinking about an artificial category label that expresses a 'forensically correct' determination (given the quality and quantity of available information) that, at best, they can only divine. It encourages group-thinking rather than rigorous, self-critical assessment in the light of the individual expert's actual knowledge and proficiency. This perspective goes fundamentally against recent developments in forensic science [e.g., 22] that precisely aim at moving away from depicting forensic science proficiency as a mystic skill that only a restricted circle of initiated individuals possess.

What is more, as already noted by Weller and Morris [44], Dror and Scurich's extended account of the notion of error can lead to the paradoxical situation in which examiners who render factually correct conclusions (in the traditional sense of congruence with ground truth), such as an identification, would be considered to be wrong when the consensus or majority *opinion* is that the conclusion *should* be "inconclusive". Rather astonishingly, Dror and Scurich respond to this paradox by doubling down on their view according to which "it is *irrelevant* whether that determination [i.e. an individual examiner's conclusion] was 'factually correct' because the issue is about whether there is a justified basis for drawing the conclusion" [19, at p.703; emphasis added]. Clearly, Dror and Scurich's merely restate their extended definition of error outlined above, which offers no answer to Weller and Morris' valid point. One does not solve the paradox of labelling factually correct conclusions as errors by maintaining that factual accuracy is unimportant ("irrelevant"). Dismissing ground truth as irrelevant is an (almost postmodernist) opinion, not an argument. It is an opinion because it offers no reason why we should use expert *judgment* as the reference point in controlled studies when we can actually have ground truth. Thus far, it seems, discussants are talking past each other. We will further dissect this impasse in the next section.

2.5. The consequences of giving up ground truth as the reference point

At first sight, Dror and Scurich's extensive account of the notion of error provides a way to overcome an inconvenience associated with the treatment of "inconclusives" in the conventional error (rate) paradigm anchored on ground truth. The inconvenience, here, is the distortion of error rates. This is a legitimate concern and it is a laudable goal to seek ways to remedy the current situation. However, this does not exempt us from assessing the extent to which their suggested corrective implies new problems that, overall, would undermine the improvement they seek to achieve. As we have seen in the previous section, a key element of their account is an extended definition of error that allows them – under certain circumstances – to score "inconclusives" as errors. This extended definition, we remind, sets forth category labels *decided* by consensus of independent experts or majority votes by study participants as the relevant reference point against which individual examiners' conclusions are to be compared. This results in giving up on the very notion of ground truth as the relevant reference point. We shall now take a closer look at what this practically means.

First and foremost, Dror and Scurich's account requires one to assign a label to each comparison (i.e., a pair of test items) in an experiment under controlled conditions (i.e., error rate study). That label – to paraphrase Dror and Scurich – is *not* the ground truth label, but the *justifiable conclusion* given the quantity and quality of information available in the evidence. Take, as an example, a pair of test items that come from the same source: in Dror and Scurich's account, the category label for this comparison is not determined on the basis of ground truth (i.e., same source), but on whether there is sufficient retrievable quality and quantity of information in the evidence to justifiably assert "same source" (i.e., identification). That is, the comparison would be *labelled* "same source" (i.e., source identification) only if there is sufficient quality and quantity of information in the evidence. The question thus is how one is to determine such category assignments.

In all fairness, Dror and Scurich acknowledge that this is not straightforward. They note: "as a practical matter, determining which evidence falls within this category is complicated" [20, at p.335] and "determining whether evidence does – or does not – have sufficient quality and quantity of information is challenging" [19, at p.703]. Notwithstanding, they "propose two different practical and feasible ways" [20, at p.335] to assign category labels: consensus determinations by independent expert panels and majority votes by study participants. Besides discussions on how exactly to implement these proposals, on which there may be divergent views, these proposals seem pragmatic. However, pragmatism is not a sufficient criterion for judging adequacy for use in practice. It is at least equally important to ensure conceptual

[8] Medical literature also mentions the use of 2×3 tables, but in contexts where the reference method (or, standard) cannot determine the ground truth state (i.e., disease status) of a patient [31]. This is not a concern in forensic science.

[9] Note that this is an advantage over some medical applications that need to establish a target condition using a reference method or a standard.

soundness. It is here that Dror and Scurich's account encounters two problems that go to the heart of the nature and the logic of forensic inference of source.

The first problem has to do with the assumption that evidence, in virtue of its quality and quantity, *predicates* a conclusion for a given comparison. In the light of the currently best available understanding of the nature and the logic of forensic inference of source, this is simply not the case. The point has been made repeatedly, for a number of decades [41,42], and it can be reconstructed using various formal developments [e.g., 4, 5, 6]. Simply put, these developments show that evidence can, at best, be *informative* with respect to competing propositions regarding ground truth. But evidence cannot logically predicate a conclusion, whether it is a conclusion in the sense of a traditional ground truth label, or the assignment of a category label in Dror and Scurich's scheme. In other words, making a defensible conclusion in the conventional identification paradigm of "identification – exclusion – inconclusive" requires more than the data one has, and more than the hypothesis one believes to be true: it requires a value judgment [42]. Note that this is merely one among many reasons for the observable move away from the aforementioned identification paradigm. But, as mentioned at the beginning of this paper (Section 2.1), as long as there is still a widespread use of this scheme, it remains important to expose its shortcomings and those of concepts, such as Dror and Scurich's account, that build on it.

The second problem pertains to the assumption that a given comparison can only have *one* admissible conclusion. To be clear, we do not contest the logical proposition that the compared items either come or do not come from the same source. To assert this is tautological, for only one of the two ground truths can be true. But this does not imply that a given comparison necessarily has only one associated conclusion of either "identification", "exclusion" or "inconclusive". Arguing otherwise would mean to confuse item category labels with response categories, which is equivalent to conflating possible states of the world with (probabilistic) statements about the world. Again, there are multiple reasons for making this distinction, and we are only briefly going to mention a few of them. Firstly, for a given comparison, the perceived quality and quantity of information – the so-called findings or observations – inevitably vary between examiners, as a function of their (background) knowledge, experience, and proficiency. Secondly, even for a fixed finding or observation, there is not a single probative value to be assigned. For example, in the field of friction ridge (i.e., fingermark/-print) examination, a variety of (probabilistic) models exist, relying upon different sources of data [e.g., 33, 43], and this can lead to differences in output. Third, even if there were a single (or agreed) probative value for a given piece of evidence (or configuration of features), this would still not entail that there is only *one* possible conclusion (i.e., examiner's response) simply because the various necessary ingredients (e.g., utilities/losses, prior information) for *deciding* logically upon a conclusion may assume different configurations. This precludes a univocal – let alone prescriptive – conclusion. Finally, even if scientists endeavoured to instantiate all the necessary logical ingredients for deciding upon a conclusion, they would find themselves in a stalemate insofar as some of the components, especially value judgments, go beyond the scope of their expertise. Thus, persisting in their endeavour would come at the price of becoming unscientific [41,42].

In view of these observations, we cannot subscribe to Dror and Scurich's assertion that "establishing that inconclusive decisions can be errors is theoretically and conceptually justified and clear" [20, at p. 335]. For the same reasons, we can also not subscribe to Dror and Scurich's contention that "if some examiners conclude an identification (or exclusion) whereas other examiners conclude an inconclusive, then at least some of the examiners are mistaken" [20, at p.334].

2.6. A closer look at a proposal for scoring "inconclusives" for the purpose of error rate studies

Dror and Scurich call the traditional, reality-based error rate study design "misleading" [20, at p.335]. In what they refer to as the "suggested and correct study design" [20, at p.335], they propose to replace the two ground truth labels "same source" and "different source" by three category labels. These do not reflect mutually exclusive and exhaustive states of the world (ground truth), but the *ascertainable conclusions* "same source", "different source" and "inconclusive", which are epistemic in nature. As explained in the previous sections, these states are the *justifiable* conclusions, to be determined either through a consensus or a majority vote. Proceeding in this way, so the idea, one can score an "inconclusive" decision as erroneous whenever the test items at hand are labelled either as *ascertainable* "same source" or "different source". Provided that one can sort out and agree on a way to ascertain the category label for each test pair, this modified testing regime seems to offer a pragmatic way to score "inconclusives" as potential errors. Yet, there are further problems that add to the difficulties we have exposed so far. The problems have both a terminological and conceptual side, and pertain to the categories and their labels.

In Dror and Scurich's nomenclature, the category labels are referred to as what "the evidence is" [20, at p.335] in the view of human experts: i.e., an asserted evidential type. In their scheme, there are test items *declared* as same source evidence, different source evidence and "inconclusive" evidence, each representing a category label. This is terminologically confusing because category labels denote an *actual property* of test items, not an asserted evidential type. The actual problem is the other way round: evidence is the starting point, it is what needs to be assessed in the process of inference and decision-making about category membership (here: ground truth). Of course, it is possible to deliberately make up categories based on observable features, which is a common procedure in taxonomy (or, classification). However, this cannot serve as an analogy because inference of source is not taxonomy – in the same way that an arrow is not a destination.

In view of the above analysis, let us now take a closer look at the category (label) "inconclusive". It is important to remember that asserted "inconclusives" necessarily are – in terms of ground truth – either same or different source test items, regardless of whatever other labels they have been assigned. Hence, "inconclusiveness", in this sense, does not represent a ground truth state, in the same way that being unsure about how to get to London does not mean that London's existence is uncertain. This is not a rhetorically empty shell. Dror and Scurich give up on the law of excluded middle [3], i.e. on the metaphysical substrate of scientific research, only because sometimes "determining which evidence falls within [a] category is complicated" [20, p.335].

For the purposes of comparison, consider a medical test designed to detect a medical condition or the presence of a target substance in a test item. Clearly, the medical condition or target substance is either present or absent [12], it is not "ascertainably present/absent". This is not meant to ignore the practical reality that test items may be challenging and difficult to assess. For example, forensic scientists may choose to work with test items called "close non-matches" [28]. Such test items share similarities, but they are known to come from different sources. Hence, the ground truth state remains key to defining the categories to which the test items belong.

The reason why we insist on these terminological distinctions is that they have far-reaching conceptual implications that merit

clarification. As shown already, redefining categories in the way suggested by Dror and Scurich amounts to (reverse) engineer rules so as to accommodate the starting assumption that "inconclusives" are potential errors. As noted above, this differs from the traditional study design that takes the measurement of diagnosticity as the starting point, and encounters the question of how to score "inconclusives" only as a subsequent, collateral complication.

Notwithstanding, let us suppose for a moment that we adopted Dror and Scurich's conceptual scheme. Would this mean that the problem of scoring "inconclusives" has successfully been solved, and hence that the traditional and allegedly deficient error rate study designs have been fixed? The answer is: No. In fact, the contrary is the case because the error rates resulting from this modified study design no longer reflect the understanding of error of the conventional study design. The error rates in the modified study design become self-referential since they no longer refer to features of the target system.

To understand why this is so, it is again useful to recall the general (medical) testing design and ask what exactly is to be assessed by such testing. Clearly, the focus is on the *diagnostic* capacity of a test. That is, it is of interest to evaluate the performance of a test (or testing device) to provide us with information about ground truth. We are *not* interested in the test's capacity to tell us something about an artificial categorisation that may or may not be congruent with ground truth. Crucially, we cannot measure the test's diagnostic capacity when the true state of the target system is not known.[10] Similarly, in forensic science, the problem encountered and to be solved by recipients of expert information is whether the test items come from the same or from different sources (i.e., ground truth). Here, "inconclusive" does clearly not reflect a contested state of affairs — instead, it is a response category. Stated otherwise, when an expert is adduced to *help* with discriminating between the contested versions defended by parties at trial, the expert's diagnostic capacity needs to properly refer to ground truth — because this is what end-users of expert information are interested in — regardless of the chosen reporting format, i.e. the tripartite classic response scheme or an expression of the probative value of the evidence. This means that one needs information about the expert's performance in "detecting" ground truth, not an artificial categorisation of the kind supposed in Dror and Scurich's modified study design.

To reinforce the above point on the necessity of ground truth as a reference point, consider what is typically the main concern in the instant case: the concern is not *any* type of error, but — most importantly — whether the expert has ever reported "identification" decisions when *in reality* the test items came from different sources. This calls for genuine false positives anchored on ground truth. This can, first of all, be elicited as a raw count (though, as further elaborated below, not in isolation), thus avoiding any hassle over the concept of rates and the multiple types of errors that their various definitions may aggregate. Indeed, it would be shortsighted to dismiss genuine false positives as determined by the conventional study design simply because of disagreements over how to compute an error rate. What is more, regardless of the outcome of the controversy over the "correct" way to compute error rates, eliciting genuine false positives and spending efforts on scrutinising such data might turn out to be more fruitful than fighting a battle over how to compute error rates that, by definition, are typically uninformative about the instant case. The following examples illustrate the host of insights that may be drawn from inspecting genuine false positive counts: when and how errors have been made (e.g., how they distribute over time; periodically and/or in clusters; do they increase/decrease), under what kinds of circumstances (e.g., training, regular proficiency testing, close non-match studies [28], casework), the number and type of controlled studies completed by the expert, etc. Similar investigations can be undertaken for "inconclusives", in order to learn more about the conditions under which they have been provided.

It is worth noting that the medical literature on the design of and reporting on diagnostic accuracy studies is broadly in line with this perspective. For example, regarding "inconclusives" (also sometimes called "indeterminate" or "intermediate" results [e.g., 29]), it is considered good practice to report the occurrence (frequency) of this type of response, and the circumstances under which this type of response was encountered. These data provide an indication of the feasibility of a test which, in turn, informs about practical usefulness.[11] As strategies to summarise study results, medical literature commonly invokes two statistics, among others. One is the so-called test yield [40]. This is the percentage of all test results used for calculating the conventional (i.e. binary) summary statistics. The lower this percentage, the more "inconclusive" results were recorded in the study. Another figure is the so-called "effectiveness" [36]. This is the proportion of correct responses among the total number of responses (including "inconclusives"). Note, however, that test yield and effectiveness are *additional* statistics for characterising a given test as a whole, different from the sensitivity and specificity.

Turning back to the forensic science context, consider one last important consequence of digressing from the traditional study design based on ground truth. Recall that in Dror and Scurich's modified error rate study design, the conclusion "identification" for test items from the same source labelled as "inconclusive" is counted as an error. We do not contest that this has the merit of penalising reckless or overconfident "identification" responses. But we need to recognise that this comes at the price of penalising the truly competent and outperforming examiners in cases where they *correctly* identify (in terms of correspondence with ground truth) *despite* the fact that the consensus or majority vote is that the test items are "inconclusive". This means to a priori exclude dissent without even considering the possibility that these conclusions could have a defensible basis (e.g., in terms of the distinct configuration of recorded features and the probative value assigned to those features).

As noted earlier, such a testing regime would set false incentives: it would direct examiners towards divining what the mythical forensic wisdom of the consensus opinion might be (and hence enshrine the false belief in the existence of such wisdom), rather than ground truth. To the extent that consensus and majority votes imply a ceiling towards an average, the modified study design thus discourages excellence and penalises genuinely outperforming examiners. None of these incentives appear helpful for the way forward of forensic science.

What is more, the two proposed options to determine category labels in the modified study design, i.e. a panel of experts and the majority vote from study participants, have opinionism as a common denominator. According to this proposal, thus, the remedy for allegedly deficient ways to deal with "inconclusives" are panels of experts looking for points of general acceptance. But this would bring us back to the notion of "general acceptance", i.e. the test introduced by *Frye* and abolished by *Daubert* (see Section 1). Thus, reducing the scientific character of practitioners' contentions to "general acceptance" amounts to masquerading, and reenacting the *Frye*-test as a forensic method without having to argue for it.

[10] We leave aside here various ways to "establish" ground truth by proxy, such as standards or reference methods.

[11] This connects back to our argument in Section 2.3 according to which examiners are less helpful, the more often they report "inconclusive".

3. The problems of understanding expert conclusions as decisions

The mainstream controversy over how to score "inconclusive decisions", as we see it, is largely motivated by possible distortions in traditional error rate determinations, rather than derived from a rigorous conceptualisation of the problem in the first place (i.e., diagnosticity). Clearly, the quality of error rates is a valid concern, but the conceptual soundness of any proposed remedy is equally important. It is at this point that an even deeper problem looms. It relates to the decisional nature of "inconclusives" and widely escapes current debates over how to understand "inconclusives". In this section, we provide a brief sketch of the problem on two levels of analysis: the understanding of conclusions in decision-analytic terms (Section 3.1), and the notion of decisional rights in a legal perspective (Section 3.2).

3.1. The misconceived decisional nature of expert conclusions

It has become fashionable in forensic science to refer to examiners' responses (and conclusions) as decisions, but the field has struggled to use this term properly, not only in terms of understanding the conceptual implications of treating conclusions as decisions, but also in terms of aligning its practice accordingly [14,15]. The consequence of this are unwarranted and contradictory treatments of "inconclusives".

Start by taking a look at a commonly evoked line of argument: suppose an examiner concludes (or reports) "identification" when in fact the compared (test) items come from the same source. Here, "identification" is the decision and "same source" is the true state of nature. By definition, the combination of a decision and a state of nature leads to a decision *consequence* (or outcome). Here, the decision consequence is an accurate identification and is commonly called a good[12] *outcome*. It is also commonly said that the examiner made a good, right or correct *decision*. But does this really make sense? A decision-theoretic perspective reveals several confusions:

- First, the accuracy of outcomes is not equivalent to their value, i.e. (un-)desirability. Saying that a decision outcome is accurate is merely a descriptive statement that refers to the congruence with respect to ground truth. To make a statement about the "goodness", attractiveness or desirability of a decision consequence, one needs to express its value on a particular scale. Stated otherwise, we need to distinguish between what a decision consequence *is*, descriptively, from what this consequence represents to us in terms of value. In decision theory, the value of decision consequences is specified in terms of utilities or losses.
- Second, and more fundamentally, the value of a decision *consequence* must not be confused with the merit, value or "goodness" [7] of a decision. Simply put, and contrary to usage in informal conversation, the mere fact that a decision led to a desirable (or, good) outcome does not mean that a good decision was made [25]. It suffices to imagine a situation in which a consequence of high desirability had a small probability of occurrence, while the alternative decision consequence – with high probability – would have represented a profoundly undesirable result. Thus, the "goodness" of decisions cannot be equated with the values assigned to decision consequences. Instead, decision theory tells us, the criterion for scoring decisions is a function of the value of decision consequences paired with their respective probabilities of occurrence [e.g., 24, 30].
- Third, the consequence of a decision, and hence its value, is known[13] – though not always – only *after* a decision is made. Thus, rating decisions *post hoc* based on their outcomes is shallow: nothing is to be learned about the quality of a decision by inspecting its outcome [24]. The real problem, from the viewpoint of examiners, is *ex ante*: they need a way to score and compare decisions *without* being able to know how the decisions will turn out [7]. The philosopher Wittgenstein has spelled this insight out: "If there were a verb meaning 'to believe falsely' it would not have a meaningful first person present indicative." [46] It is not very helpful, therefore, to assess decisions against metaphysical concepts such as ground truths. But we need the latter for the very term true and false to make sense. This brings us back to the previous point on working out a measure for the merit of a decision based on value judgments for decision consequences and associated uncertainties. In this sense, then, a good decision is one that is made logically, acknowledging the alternatives from which to choose, the decision maker's preferences among decision consequences and the uncertainties about the real states of the world, assessed coherently in the light of inevitably imperfect data at our disposal.

The above concepts allow us to critically review the controversy over the treatment of "inconclusive" *decisions* both in general, but also in error rate studies, and to substantiate the arguments we have presented throughout this paper. We can consider the following three main points.

First, to ensure conceptual rigour, it is necessary to distinguish between "inconclusive" as a *decision* and the *consequence* of deciding "inconclusive" when the compared (test) items come or do not come from the same source. In particular, features of the latter, decision outcomes, cannot directly be carried over to statements about the quality of decisions. Thus, studies under controlled conditions (i.e., with known ground truth) primarily deal with scoring decision *consequences*, not decisions themselves.

Second, to answer the question of how a *consequence* of deciding "inconclusive" compares to the consequences of the rival decisions "identification" or "exclusion", which amounts to an expression of our preferences, we need to express the desirability of outcomes on an agreed scale (e.g., utilities or losses) that applies equally to the consequences of all decisions (i.e., the space of decisions containing the elements "identification", "exclusion" and "inconclusive"). Thus, attempting to label the consequences of "inconclusive" decisions as correct (accurate) or incorrect (erroneous) is a contradiction in terms when accuracy denotes congruence with ground truth. "Inconclusives" explicitly make no statement about ground truth. At best, we may express the value that the consequence of an "inconclusive" decision represents to us, but the descriptive vocabulary of "error" is unsuitable for this. Instead, as noted above, a measure of the value of a decision consequence is needed. Pursuing this in a full decision-theoretic analysis is beyond the scope of this paper. We solely note that such an analysis demonstrates that the consequences of "inconclusive" decisions actually have a high utility, *as long as* one is averse to false positives [8]. Hence, a negative view on the consequences of "inconclusive" decisions by

[12] We use here the term "good" informally as a synonym for "desirable". A more formal treatment of the notion of desirability uses the concept of utility.

[13] In proficiency testing, where ground truth is known for each pair of test items, the accuracy of outcomes can be known. This is different for casework, and legal decision-making in general, where ground truth is, in principle, not known and remains unknown.

default is unwarranted [8], and contrary to what is suggested by attempts to score "inconclusives" as errors.

Third, and related to the previous point, given that the *consequence* of an "inconclusive" decision may enjoy a high utility, an "inconclusive" *decision* may, in turn, be a viable competitor to "identification" and "exclusion" decisions. Hence, it is as much unwarranted to consider the consequences of "inconclusive" decisions as inadequate by default as it is unwarranted to consider "inconclusive" *decisions* as unsuitable by default.

3.2. Decisional rights

The multiple conceptual problems that affect attempts to model expert conclusions as decisions, exposed throughout the previous sections, should be reason enough to recognise the missing foundations for decision-making by experts. Yet, as we have mentioned in Section 2.1, we are aware that current forensic practice widely persists in their conventional reporting schemes. This is all the more surprising in that the reporting scheme also conflicts with legal principles, in particular with decisional rights. It is of value, thus, to briefly address this point through elements of a doctrinal analysis. It reveals the controversy over how to score "inconclusives" and the tripartite classic response categories as misguided, which should prevent us from using these notions in the first place.

To clarify the legal standpoint, we first need to revisit principles in the law of evidence, or at least the common doctrinal denominator that underpins the law of evidence in common law jurisdictions.[14] The need for interpretation of forensic evidence invokes a fundamental feature of the law of evidence, i.e. the opinion rule.[15] According to traditional evidence law doctrine, experts will present fact-finders with physical rules or general principles and *help* them apply the domain-specific general knowledge to the evidence introduced in the respective case. The term 'help' is bearing here much of the conceptual burden – in multiple ways. The expert witness is bound to testify only *within* his or her area of expertise; any step outside that strictly circumscribed area constitutes a procedurally forbidden invasion of the fact-finder's province.[16] Lawton LJ's enduring dictum in *Turner*, which predicates admissibility on necessity, sets the tone in this area of English law: "If on the proven facts a judge or jury can form their own conclusions without help, then the opinion of an expert is unnecessary".[17] It is, thus, the jury that has the decision-making prerogative and any usurpation thereof is procedurally barred and, as we have noted previously (Section 2.5), scientifically unwarranted.

Within the exclusionary scope of the opinion rule fall expert witnesses who express unsolicited opinions, speculate or make value judgments about the facts of the case. This includes source attribution determinations (SADs) that pre-empt the decisions of the fact-finder. Only jurors or triers of fact more generally are democratically legitimised and procedurally authorised to make judgments under uncertainty, resolving thus factual disputes about unique historical events. Usurping the territory of triers of fact by taking over decisional rights which forensic experts do not have, nor should have, triggers the odd jurisprudential reprimand. As the Court of Appeal (England and Wales) reiterated in *Brennan*[18] "in criminal trials cases are decided by juries, not by experts". Making *decisions* about ultimate issues, including intermediate issues such as SADs, which require value judgments, is not part of forensic practitioners' duties.

The normative structure of the criminal process should not be ignored as it may help prevent dubious, indeed nonsensical results in forensic science practice. It is important to realise that the criminal process is not a ritual with a recurrent and foreseeable outcome. Criminal processes have open outcomes. On the flipside, criminal process officials do not make the rules as they go along. Legal adjudication, especially in the criminal context, is not an anything-goes activity. The defendant has a default status – presumed innocence – which he shall retain unless and until his guilt has been proven to the requisite standard of proof. In that sense, there can be no such thing as an "inconclusive decision" in the criminal process.

One might think that the "decisions" which are widely referred to in discussions surrounding the reporting practice of forensic scientists are of a different type. That is, so the argument, the scientist's decisions (i.e., reported conclusion) do not directly affect the defendant/suspect's status; instead, these decisions refer only to the source of the forensic trace item and, more specifically, to the question of whether it comes from the same source as the reference item, or from a different source. This line of defence, however, is weak in view of the doctrinal framework that validates and restricts the operations of forensic practitioners. In this sense, there is no such notion as "inconclusive decisions" – forensic scientists do not decide anything and they certainly do not, indeed cannot provide SADs. Expert witnesses and, more generally, forensic practitioners lack the necessary decisional rights.

4. Conclusions

In all, our analysis does not leave much intact from recent attempts to label "inconclusives" as errors. To be clear, we do not argue that we should not focus on errors or error rates. Quite the contrary: our point is that recording errors is important, but errors of the suitable kind, and task-specific information regarding such errors. By this we mean *genuine* errors determined with respect to ground truth, rather than with respect to artificial category labels which lack a coherent conceptual basis and, thereby, lead to paradoxes in practice (e.g., penalising factually correct responses).

Similarly, arguing against the scoring of "inconclusives" as potential errors does not mean to dismiss the focus on "inconclusive" decisions. They are an important category of decisions precisely because they allow "identifications" and "exclusions" to be used only when stringent requisite conditions are satisfied, not least because we value adverse outcomes of decisions such as "identification", i.e. false positives, as highly undesirable. This assertion is not based on mere intuition, but can be demonstrated through a decision-theoretic analysis [8].

It is also important, of course, to monitor the extent to which examiners give "inconclusives". Clearly, the excessive use of "inconclusives" is self-defeating, thus questioning the need to resort to a conceptually flawed scoring scheme to expose the allegedly erroneous nature of "inconclusives". The proportion of "inconclusives" among the total number of responses is a simple descriptive measure that provides an indication of the responsiveness, or assertiveness, of an examiner, but inquiries should not stop there. It is equally important to consider the number of

[14] Any reference to the common law of evidence is, to a certain extent, an oversimplification, but acceptable for the purposes of this paper. Our account here covers mainly the law of evidence in England and Wales.

[15] Already since the late 18th century, the common law has erected a general ban on opinion evidence in order to entrench the decision-making prerogative of the fact-finder. The early leading case is *Folkes v Chadd* (1782) 3 Doug K.B. 157. See *Robb* (1991) 93 Cr. App. R. 161 CA (E + W).

[16] *R v. Davies* [1962] 3 All E.R. 97.

[17] *Turner* [1975] Q.B. 834 at 841 CA.

[18] *R v. Brennan* [2015] 1 WLR 2060, [2014] EWCA Crim 2387, [43].

"inconclusives" given by an examiner in the context of the testing conditions. It makes a difference, for example, whether an examiner performed comparisons in a reputedly easy proficiency test, or whether the test items were close non-matches. What these considerations show is that the problem of expert proficiency is in many ways more subtle than what a summary in terms of a crude statistic, such as a rate, could provide. Thus, instead of dismissing the traditional diagnostic performance assessment matrix (rooted in ground truth) in a confusing debate over how to summarise test responses, it seems more worthwhile to deploy efforts to clarify the nature of the problem of diagnostic capacity in the first place, and to scrutinise task-specific data regarding the conditions under which *genuine* errors have been committed. This should prove insightful for both recipients of expert information and experts themselves. Similar viewpoints can be found in the vast medical literature on diagnostic accuracy studies [e.g., 12, 29, 39].

Besides these practical conclusions, attempts to frame "inconclusives" as potential errors also suffer from conceptual problems. In particular, giving up on ground truth by introducing an extended, tripartite classification scheme, interferes with the metaphysical substrate that underpins meaningful scientific research. The very idea of empirical research presupposes the intelligibility of mutually exclusive and exhaustive propositions. Whatever our degree of belief is, a forensic item or trace either has or does not have a given person of interest as its source: *Tertium not datur* ("no third possibility being given") – Aristotle's principle of the excluded middle [3]. Erecting "inconclusive" to a classification category thus means to chase philosophical shadows. It amounts to confusing binary philosophical concepts (correspondence to the true facts as our metaphysical substrate) with questions that are epistemic and multivalued in nature. Rejecting the argumentative implications of the excluded middle means to break with millennia of philosophical tradition, and to get trapped unknowingly inside a philosophical "bottle". Our aim should be, in Wittgenstein's words, to "show the fly the way out of the fly-bottle" [46, para.309].

Acknowledgements

This research was supported by the *Swiss National Science Foundation* (No. BSSGI0_155809), the *Fondation pour l'Université de Lausanne* and the *Société Académique Vaudoise*.

Declaration of competing interest

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

References

[1] J.-P. Allery, N. Telmon, A. Blanc, R. Mieusset, D. Rougé, Rapid detection of sperm: comparison of two methods, J. Clin. Forensic Med. 10 (5–7) (2003).
[2] Aristotle. *Metaphysics (Hippocrates G. Apostle*, trans.), first ed., Indiana University Press, Bloomington, 1966.
[3] Aristotle, in: Metaphysics: Books Gamma, Delta, and Epsilon (Christopher Kirwan, ed.), second ed., Clarendon, Oxford, 1993.
[4] A. Biedermann, J. Vuille, Understanding the Logic of Forensic Identification Decisions (Without Numbers), sui-generis, 2018, pp. 397–413.
[5] A. Biedermann, S. Bozza, F. Taroni, Decision theoretic properties of forensic identification: underlying logic and argumentative implications, Forensic Sci. Int. 177 (2008) 120–132.
[6] A. Biedermann, S. Bozza, F. Taroni, The decisionalization of individualization, Forensic Sci. Int. 266 (2016) 29–38.
[7] A. Biedermann, S. Bozza, F. Taroni, P. Garbolino, A formal approach to qualifying and quantifying the 'goodness' of forensic identification decisions, Law, Probability and Risk 17 (2018) 295–310.
[8] A. Biedermann, S. Bozza, F. Taroni, J. Vuille, Are inconclusive decisions in forensic science as deficient as they are said to be? Front. Psychol. 10 (2019) 1–9.
[9] A. Biedermann, S. Bozza, F. Taroni, J. Vuille, Letter to the Editor – commentary on: Dror IG, Langenburg G. "Cannot decide": the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide, J. Forensic Sci. 64 (2019) 318–321.
[10] C. Champod, C. Lennard, P. Margot, M. Stoilovic, Fingerprints and Other Ridge Skin Impressions, second ed., CRC Press, Boca Raton, 2016.
[11] C. Champod, H. Eldridge, S. Lambert, A primer on error rates in fingerprint examination, Available online at: https://doi.org/10.5281/zenodo.3734560, 2020.
[12] J.F. Cohen, D.A. Korevaar, D.G. Altman, D.E. Bruns, C.A. Gatsonis, L. Hooft, L. Irwig, D. Levine, J.B. Reitsma, H.C. W de Vet, P.M. M Bossuyt, STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration, BMJ Open 6 (2016) 1–17.
[13] S.A. Cole, More than zero: accounting for error in latent fingerprint identification, J. Crim. Law Criminol. 95 (2005) 985–1078.
[14] S.A. Cole, Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States, Law, Probability and Risk 13 (2014) 117–150.
[15] S.A. Cole, A. Biedermann, How can a forensic result be a "decision"? A critical analysis of ongoing reforms of forensic reporting formats for federal examiners, Houst. Law Rev. 57 (2020) 551–592.
[16] S.A. Cole, B.C. Scheck, Fingerprints and miscarriages of justice: "other" types of error and a post-conviction right to database searching, Albany Law Rev. 81 (2018) 807–850.
[17] U.S. Department of Justice, Letter to judge Patrick Schiltz, Chair of the advisory committee on the federal rules of evidence, Available online at: https://www.uscourts.gov/sites/default/files/20-ev-aa_suggestion_from_doj_-_rule_702_0.pdf, 2020.
[18] U.S. Department of Justice, Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline, 2020 vers. 8.15.20. Available online at: https://www.justice.gov/olp/page/file/1284786/download.
[19] I. Dror, N. Scurich, Continued confusion about inconclusives and error rates: reply to Weller and Morris, Forensic Sci. Int.: Synergy 2 (2020) 703–704.
[20] I.E. Dror, N. Scurich, (Mis)use of scientific measurements in forensic science, Forensic Sci. Int.: Synergy 2 (2020) 333–338.
[21] H. Eldridge, M. De Donno, C. Champod, Testing the accuracy and reliability of palmar friction ridge comparisons – a black box study, Forensic Sci. Int. 318 (2021) 110457.
[22] I.W. Evett, The logical foundations of forensic science: towards reliable knowledge, Philos. Trans. R. Soc. Lond. B Biol. Sci. 370 (1–10) (2015).
[23] A.R. Feinstein, The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions, J. Clin. Epidemiol. 43 (1990) 109–113.
[24] R.A. Howard, A.E. Abbas, Foundations of Decision Analysis, Pearson, Essex, 2016.
[25] R.A. Howard, Decision analysis: applied decision theory, in: D.B. Hertz, J. Mélèse (Eds.), Proceedings of the Fourth International Conference on Operational Research, Wiley-Interscience, New York, 1969, pp. 55–71.
[26] D.H. Kaye, The validity of tests: caveant omnes, Jurimetrics Journal 27 (1987) 349–361.
[27] J.J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter, Hastings Law J. 59 (2008) 1077–1100.
[28] J.J. Koehler, S. Liu, Fingerprint error rate on close non-matches, J. Forensic Sci. 66 (2021) 129–134.
[29] J.A. Landsheer, The clinical relevance of methods for handling inconclusive medical test results: quantification of uncertainty in medical decision-making and screening, Diagnostics 8 (2018) 1–11.
[30] D.V. Lindley, Making Decisions, second ed., John Wiley & Sons, Chichester, 1985.
[31] D.B. Matchar, D.L. Simel, J.F. Geweke, J.R. Feussner, A Bayesian method for evaluating medical test operating characteristics when some patients' conditions fail to be diagnosed by the reference standard, Med. Decis. Making 10 (1990) 102–111.
[32] J.L. Mnookin, The uncertain future of forensic science, Daedalus 147 (2018) 99–118.
[33] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a fingerprint comparison: a new paradigm, J. Roy. Stat. Soc. 175 (2012) 371–416.
[34] Organization of scientific area committees for forensic science (OSAC) friction ridge subcommittee, Standard for friction ridge examination conclusions (2018) vers. 1.0.
[35] P.C.A.S.T. President's, Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President, Washington, D.C., 2016.
[36] T. Poynard, J.-C. Chaput, J.-P. Etienne, Relations between effectiveness of a diagnostic test, prevalence of the disease, and percentages of uninterpretable results: an example in the diagnosis of jaundice, Med. Decis. Making 2 (1982) 285–297.
[37] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int. 230 (2013) 156–169.
[38] M.J. Saks, J.J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (2005) 892–895.
[39] B. Shinkins, M. Thompson, S. Mallett, R. Perera, Diagnostic accuracy studies: how to report and analyse inconclusive test results, BMJ 346 (2013) f2778.
[40] D.L. Simel, J.R. Feussner, E.R. Delong, D.B. Matchar, Intermediate, indeterminate, and uninterpretable diagnostic test results, Med. Decis. Making 7 (1987)

107–114.
[41] D.A. Stoney, What made us ever think we could individualize using statistics? J. Forensic Sci. Soc. 31 (1991) 197–199.
[42] D.A. Stoney, Discussion on the paper by Neumann, Evett and Skerrett, J. Roy. Stat. Soc. 175 (2012) 399–400.
[43] H.J. Swofford, A.J. Koertner, F. Zemp, M. Ausdemore, A. Liu, M.J. Salyards, A method for the statistical interpretation of friction ridge skin impression evidence: method development and validation, Forensic Sci. Int. 287 (2018) 113–126.
[44] T.J. Weller, M.D. Morris, Commentary on: I. Dror, N. Scurich "(Mis)use of scientific measurements in forensic science" Forensic Science International: Synergy 2020, Forensic Sci. Int.: Synergy 2 (2020) 701–702, https://doi.org/10.1016/j.fsisyn.2020.08.006.
[45] S.M. Willis, L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, A. Nordgaard, C.E.H. Berger, M.J. Sjerps, G Zadora Jj Lucena-Molina, C.C.G. Aitken, T. Lovelock, L. Lunt, C. Champod, A. Biedermann, T.N. Hicks, F. Taroni, ENFSI Guideline for Evaluative Reporting in Forensic Science, Strengthening the Evaluation of Forensic Results across Europe (STEOFRAE), 2015. Dublin.
[46] L. Wittgenstein, in: Philosophical Investigations (ed. by P.M.S. Hacker and J. Schulte), fourth ed., Wiley-Blackwell, Chichester, 2009.