Forensic Science International: Synergy 4 (2022) 100222

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

The strange persistence of (source) "identification" claims in forensic literature through descriptivism, diagnosticism and machinism

Alex Biedermann

University of Lausanne, School of Criminal Justice, 1015 Lausanne-Dorigny, Switzerland

ARTICLE INFO

Keywords:
Identification/individualisation
Descriptive research
Diagnostic reasoning
Machine learning
Artificial intelligence

ABSTRACT

Many forensic scientists consider that identification (individualisation) – in the sense of statements of the kind "the questioned item and the known item come from the same source" – is a concept that is central to their discipline. This is so despite decade-long, fundamental critiques levelled by both practitioners and academics against the conceptual and practical feasibility of forensic identification. Oddly, there is a constant stream of publications in (peer-reviewed) forensic science journals that treat forensic identification axiomatically as a valid object of study, sidestepping the fundamental critiques. This paper reviews and discusses three exemplary strands of publications that exemplify this persistent trend. These strands are called descriptivism, diagnosticism and machinism. The latter term refers to methods borrowed from the now increasingly popular approaches used in the field of machine learning. In turn, descriptivism and diagnosticism refer to general design aspects of mainstream research methods, illustrated here through a critical review of two recent papers on, respectively, forensic odontology and a framework for interpreting fingerprint evidence. The critique of the use of 'identification' in these strands of publication includes, but goes beyond, semantic details and the reiteration of long-known shortcomings of obsolete technical language such as 'match' and 'matching'. Specifically, this paper exposes deeper problems such as the subtle and argumentatively unfounded carrying-over of source conclusions to ultimate issues and the use probability concepts for questions that require more than the mere quantification of uncertainty. This paper submits that in order to foster trust in an era of continually expanding publishing activities, it should be a vital interest to forensic science journals to better examine what identification-related research can and cannot legitimately purport to achieve.

1. S.A.D. but true

One of the most striking elements in the recent roadmap – or "vision" [41] – issued by the European Network of Forensic Science Institutes (ENFSI)[1] is the persistence of the concept and practice of 'individualisation', i.e. the (testimonial) claim that the potential donor pool of a forensic trace can be reduced to a single source. The document asserts, for example, that biometrics "allows a person to be *individualised* and authenticated, based on a set of recognisable and verifiable data, which are very distinctive." [41, at p. 2, *emphasis added*].[2] The idea that one can apply naked statistics to the individual is not insignificant or inconsequential. As the ENFSI perspective makes plain, "pattern recognition of features of comparison for individualisation and source attributions" – widely known as S.A.D. (Source Attribution Determination) – is still to be counted among the "fundamentals in forensic science" [41, at p. 3]. The ENFSI thus reiterates the view, widely attributed Kirk [64], that individualisation is the essence of forensic science, and illustrates its ubiquitous character. Yet, individualisation and, more generally, the application of naked statistical evidence to individuals is in conflict with the readiness of courts to (i) sanction transgressions of boundaries of (expert) competence and (ii) reinforce the requirement of 'specific evidence'. This raises the question of whether forensic science has grasped

E-mail address: alex.biedermann@unil.ch.

[1] The European Network of Forensic Science Institutes comprises more than seventy forensic institutes from European countries (including the U.K.), whose overarching goal is to "ensure that the quality, development and delivery of forensic science throughout Europe is at the forefront of the world" [41, at p. 1].

[2] It is not contested in this paper that *verification* in the sense of a one-to-one comparison based on good quality input and reference data, e.g. in the context of access control, is operationally feasible and widely practiced. What is contested here is the general claim of inference of source where a forensic trace of unknown source and variable quality is compared to many potential sources. By way of example, consider that "very distinctive" [41, at p. 2] data are simply insufficient to justifiably reach an individualisation (see also Section 2.1 for further discussion).

https://doi.org/10.1016/j.fsisyn.2022.100222
Received 23 December 2021; Received in revised form 25 January 2022; Accepted 10 February 2022

the signs of time.

During one of its symposia organised in 2017, the National Institute of Standards and Technology (NIST) issued a tweet, quoting Ian Evett as saying[3]:

"The identification paradigm is going to die, because as scientists we realize there's no basis for it." [26]

As of the time of writing of this paper (December 2021), this tweet counted only 4 likes and 2 retweets. This is a rather limited echo among the tens of thousands of NIST's followers on Twitter.[4] Yet, Evett's statement is by no means the first of its kind. In 2014, Cole [31] critically exposed the persistence of the notion of identification – or: forensic individualisation – in ongoing reforms of the reporting practice of forensic fingerprint examiners, despite the field's shaky conceptual foundations. And, some 30 years ago, Stoney asked: "What made us ever think we could individualize using statistics?" In essence, Stoney argued, trying to prove uniqueness is a "ridiculous notion" [93, at p. 198], because it is an attempt to go beyond what can be scientifically justified.

One would have hoped that scientists take calls for revising reporting practice in identification matters to heart. Indeed, in a widely cited paper published in the journal *Science*, Saks and Koehler forecasted "The coming paradigm shift in forensic identification science" [83], i.e. a move towards more robust and less categorical statements. However, as of today, any fair assessment of the current state of practice in forensic science must admit that this paradigm shift has not (yet) materialised, quite to the contrary. Several observations attest to this fact. A recent study by Swofford et al. [96] found that "almost no respondents currently report probabilistically" and "more surprisingly, most respondents who claimed to report probabilistically, in fact, do not". Even more surprisingly, this study also revealed "that two-thirds of respondents perceive probabilistic reporting as 'inappropriate'." In the same vein, one can observe that the term "identification" is still part of approved reporting language, such as the Uniform Language for Testimony and Reports (ULTR) [106], issued by United States Department of Justice (U.S. DOJ). Moreover, one of the world's oldest and largest forensic associations, the International Association for Identification,[5] even features the term prominently in its name. And yet, ongoing discussions demonstrate that current reporting formats – based on the traditional term "identification" – are a serious cause of concern for the legal system at large. An example for this is the recent approval (for publication for public comment) by the Judicial Conference Committee on Rules of Practice and Procedure of amendments to Rule 702 of the Federal Rules of Evidence (concerning expert witness testimony).[6]

Strangely, large parts of current forensic science literature acknowledge little to nothing of the selected assertions and associated arguments mentioned above. As readers can easily verify themselves, nowadays virtually every single new issue of the mainstream forensic science journals contains one or more articles in which authors describe or refer to so-called "identification" methods, that is techniques and procedures enabling scientists – so the claim – to provide conclusions about the source (or, origin) of traces, marks and impressions of unknown source. A few examples will be presented in later parts of this paper. Thus, what is at question here is how forensic science can (hope to) build a coherent body of knowledge, which is a necessary requirement for the production of defensible information in the legal process, when the notion of identification continues to be treated in such an ambiguous manner.

This paper argues that problematic identification claims in forensic science literature are largely self-perpetuating and, often, manifest themselves in disguise. They are theory-averse, paired with tendencies to misconceive professional practice as science. Three broad themes will be exposed here, i.e. publication trends, through which identification claims are commonly made. These themes are called here descriptivism, diagnosticism and machinism. The latter term is used as a placeholder for certain machine learning (ML) approaches borrowed from the field of artificial intelligence (AI). What descriptivism, diagnosticism and machinism have in common is that they neither acknowledge nor offer a solution to the conceptual impossibility of forensic identification, but take the contrary as premise. This paper submits that this is problematic in at least three ways. First, publications that handle the notion of identification in a vague and a logically unsound manner contribute to condoning unscientific attitudes and practices. Second, such publications continue to foster unrealistic expectations among consumers of forensic science services – expectations that go beyond what science on its own can provide. Third, such publications undermine both, trust in the publication business itself and progress in forensic science: because when no lessons from foundational literature are drawn, so-called 'original research papers' simply cannot be well grounded. Thus, in light of the seemingly ever-expanding publishing business, there is a greater need than ever for editors of forensic science journals to ensure (i) the proper acknowledgement of the particular conditions and circumstances under which claims of identification are and are not warranted, and (ii) a better distinction between improvements on merely analytical or tool problems as opposed to advancement on core forensic science topics, such as identification.

This paper is organised as follows. Sections 2, 3 and 4 present and discuss, respectively, the notions of descriptivism, diagnosticism and machinism as introduced above. Conclusions are presented in section 5. The three core sections 2 to 4 present self-contained discussions that may be read in any order. The arguments presented in sections 2 and 3 are illustrated through a critical review of two recent papers on, respectively, forensic odontology and a framework for interpreting fingerprint evidence.

2. Descriptivism

2.1. Preliminary remarks on forensic identification

Before looking into a first type of way in which identification claims arise in literature, descriptivism (Section 2.2), let us briefly clarify what exactly is meant here by the term "identification", and recall some of the multiple reasons why identifications rendered by practising scientists are unwarranted.

Throughout this paper, the term "identification" is used as a synonym for "individualisation", i.e. the statement that a particular object, trace or mark comes from a *specific source* (object or person). This is to be distinguished from identification understood as classification, i.e. "placing the object in a restricted class" [64, at p. 236]. This paper is also not dealing with identification in the sense of (biometric) verification, which is based on a one-to-one comparison (see e.g. Ref. [39] for further discussion of the relationship between biometrics and forensic science). Further, the use of the term identification (individualisation) here assumes what is called an open set framework, i.e. the consideration of a large pool of potential sources. In the widest sense, this amounts to the so-called "Earth population paradigm" [25]. Hence, this paper does not assume the closed set framework [25] where the number of potential sources is limited and could be examined exhaustively, a situation in which identification can sometimes be achieved by elimination (e.g., when there are clear differences in class characteristics).

The reasons why identification claims are unwarranted range,

[3] https://twitter.com/NIST/status/879711283884564481?s=20.

[4] It is acknowledged here that practitioners may be restricted in expressing views on public forums, thus limiting social media analytics as a source of information.

[5] https://www.theiai.org/.

[6] Minutes of the Committee on Rules of Practice and Procedure, June 22, 2021, p. 20. Available at: https://www.uscourts.gov/sites/default/files/2021-06-22_standing_committee_minutes_final_0.pdf (last accessed January 19th, 2022).

broadly speaking, from the less than perfect nature of real-world types of evidence and the invalid character of the utterance 'to the exclusion of all others', to the conceptual limitation according to which identifications require more than the evidence one has. Most importantly, identifications require value judgments, which is a direct consequence of understanding identifications as decisions [12,16]. This means that scientists who make identifications, accept – explicitly or not – to combine both scientific elements and assumptions about non-scientific case aspects.[7] The consequence of this has concisely been noted by Stoney: "This [i.e., making identifications] created an overwhelming and unrealistic burden, asking fingerprint examiners, in the name of science, for something that science cannot provide. As a necessary consequence, fingerprint examiners became unscientific" [95, at p. 400].

2.2. Defining descriptivism in the context of the professionalisation of forensic science (vs. the scientification of forensic practice)

Descriptivism is understood here as the observational study of the reporting schemes used by practitioners and the interpretation of the observed reporting behaviour – the *is* – as the *ought*. Stated otherwise, what is meant here are studies that provide a descriptive account of current reporting practice regarding identification and that imply that the observed reporting behaviour is the way practice *should* be. To be clear, the problem here is not descriptivism *per se*, because observation and description are important pillars of empirical science. The problem lies in the uncritical treatment of the object of study – the practice of rendering identification conclusions – as valid *despite* strong arguments to the contrary (as mentioned in Section 2.1), which contributes to perpetuating identification claims in forensic literature.

One way to illustrate descriptivism, and to recognise publications that fall prey to it, is to see descriptivism as an instance of the professionalisation of forensic science. By this is meant, more simply stated, publications that defer to forensic practice and treat it as a standard or norm.[8] As will be further elaborated below, this is problematic, especially where the practice of interest deals with identification.

The professionalisation of forensic science must be distinguished from the (desirable) counter-perspective, concerned with the scientification of forensic practice. What is meant by this is that, generally, one would hope that insights from a scientific analysis of the notion of identification would contribute to render forensic practice more scientific, and hence more defensible and trustworthy. This would be in analogy to many other sciences that help advance the respective professional fields (e.g., medicine, engineering, etc.), resulting in the improvement in matters such as of the quality of life or the protection of our environment. A forensic example for this is the revolution that fundamental research in genetics brought to the forensic practice of analysing biological traces. The reverse, however, is not obvious: could forensic practitioners "sell" ideas from the forensic profession as science, as some "peer-reviewed" publications suggest? The truth is that this already happens widely. A prime example is, as will be shown in due course, the area of forensic identification, especially publications that approach forensic identification practice *descriptively*. The question is whether one should "buy" such ideas. The next section argues for caution.

2.3. An example of descriptivism in forensic odontology

As a recent example to illustrate descriptivism in the context of forensic identification, consider the paper "Interpretation, confidence and application of the standardised terms: identified, probable, possible, exclude and insufficient in forensic odontology identification" [29]. This paper is chosen here because its title is filled with terms and concepts that are at the heart of the discussion here (for another recent example see e.g. Ref. [110]). The paper [29] investigates and reports on the use, by forensic odontologists, of conclusion scales involving varying degrees of identification when confronted with radiographs from test cases. This research is inspired by the practice of forensic odontology professionals who use standardised conclusions, such as identification, and varying degrees thereof, as part of their work. Note, however, for the reasons exposed in Section 2.1, a forensic odontologist cannot – in the same way as fingerprint examiners and other identification-"ists" – reach *scientific* identification conclusions.[9] So where lies the problem?

The crucial distinction, not made in the above-mentioned paper, is that while the observable reporting practice of forensic odontologists can be studied scientifically, the nature of examiners' conclusions is *not* scientific and will remain so, regardless of the scientific character of the descriptive method used to study it. Before elaborating further on this point, an important side-note needs to be added: it is not contested nor criticised here that there are practical situations in which identification decisions are needed and made on an operational basis. A typical example is the processing of disaster victim identification (DVI) cases, for which an internationally agreed standard-setting guide exists.[10] But identification, in this context, is part of an administrative procedure. More specifically, identification is a conclusion reached by a commission, also called "Identification board", entrusted (mandated) with this task.[11] Such a conclusion is then put forward to a relevant judicial entity for further consideration. Yet, even though forensic odontology in DVI is considered a "primary identifier" [59], along with friction ridge and DNA analysis, identification commissions make decisions based on possibly multiple reports received from different areas of expertise, not necessarily odontology alone. The focus of the paper by Chiam et al. [29], however, is not committee decisions, but individual odontologists' conclusions, which is problematic in several respects.

Firstly, reporting, without suitable caveats, on the study of odontologists' use of conclusion scales involving the term identification has the potential to suggest that closely related areas of expertise, too, could proffer identification conclusions, and varying degrees thereof. Specifically, the suggestion that odontologists can "identify" a deceased person, and that odontologists' identifications are scientific, is prone to foster the idea that bitemark examination, a sub-field of forensic odontology, can "identify" the person that left a given bitemark – to the exclusion of all others. However, this idea is not warranted, for various reasons (see Ref. [84] for a critical review). Besides the conceptual reasons for the impossibility of inference of source, mentioned in Section 2.1, bite-mark examiners typically work in criminal cases where the notion of inference of source has a meaning that is completely different from the one used in committee decisions regarding the identity of human remains in DVI work. While the starting point in the latter case is an actual human remain (teeth, jaw, etc.), the starting point in the former case is a mark supposedly left by human teeth. *If* such a mark or

[7] Using a formal analysis, it can be shown that a decision to identify rests upon a combination of probabilities, characterising uncertainty regarding the truth of the proposition of common source, and utilities, characterising the desirability of decision consequences.

[8] Left aside here is the openly unscientific point of view according to which, by definition, forensic feature comparison is experience-based, and source identification conclusions do not claim to be scientific. See e.g. the position advocated by the U.S. DOJ as reported by the Advisory Committee on Evidence Rules [1, at p. 20].

[9] It is acknowledged here that the examination and comparison of radiographs may involve science, but identification *decisions* as such are not scientific.

[10] See the INTERPOL Disaster Victim Identification (DVI) Guide [58], available at https://www.interpol.int/How-we-work/Forensics/Disaster-Victim-Identification-DVIwe-work/Forensics/Disaster-Victim-Identification-DVI (last accessed January 22nd, 2022).

[11] Likewise, medical examiners or coroners in some jurisdictions have the mandate and authority to establish the identity of a deceased.

trace is actually the consequence of biting, then, by definition, it is an incomplete and/or distorted representation of dentition, in the same way that a fingermark is an imperfect representation of friction ridge skin surface. For this reason, marks (of any kind) can exhibit similarity with reference materials (i.e., control impressions) from sources *other* than the actual source, which poses a fundamental obstacle to identification. That is, however peculiar a given human dental configuration may be, bitemarks are a necessarily imperfect representation of such features; hence, variability in dental features across humans *per se* cannot serve as a warrant for identification of the source of bitemarks. Parts of the scientific community appear to have gone some way towards acknowledging this limitation. For example, the Standards and Guidelines for Evaluating Bitemarks (vers. 2-19-2018)[12] of the American Board of Forensic Odontology (ABFO) does not caution identification conclusions: i.e., the Section "1. Standards" states that "An ABFO Diplomate shall not express conclusions unconditionally linking a bitemark to a dentition", and Section 2.c only covers the conclusions "Excluded", "Not excluded" and "Inconclusive". While these are debatable terms in their own right (their discussion is beyond the scope of this paper, but see e.g. Ref. [9]), the important point for us here is that this terminology does not contain the term "identification". Yet, in published literature, one still sees overstatements such as "Analyses based on the individual characteristics of the dentitions can identify the biter" [48, at p. 1] and bitemark study designs in which "identification" is a response type [e.g., 48]. A further example of a forensic field that is characterised by these limitations is the study of hand vein patterns (see e.g. the spurious claim that "[i]f only matches are seen, the identity of the suspect is highly plausible" [55, at p. 6]).

Secondly, the mere fact that there are commissions entrusted with disaster victim identifications through consensus decisions should not be taken as a suggestion that an individual examiner's conclusion, e.g. by a forensic odontologist, should also use terminology involving the term identification. There is no reason, in principle, why forensic odontologists could not use a reporting format that aligns with the principles of evaluative reporting applicable throughout forensic science [109], and forensic genetics in particular [51]. These principles focus on the value of the findings – similarities and differences observed during comparative examinations – given competing propositions about the source (identity) of the examined materials. The result is a measure of the extent to which the findings are capable to discriminate between propositions that specify a particular person of interest versus an unknown person and/or specific other person(s) of interest as the source. Most importantly, this reporting format abstains from opining directly on propositions, such as identity of source, and degrees to which such propositions are thought to be true.[13]

Thirdly, a descriptive study that merely summarises the responses of examiners given during test cases, using "identification language", provides no insight into the rationale of how to go from observations to conclusions. Chiam et al.'s [29] finding that examiners' confidence in their conclusions varies by case difficulty is neither surprising nor particularly helpful. The deeper question of how to logically assess the value of a finding for evaluative purposes, regardless of the quality of the evidence (i.e., the "difficulty" it poses during examination), remains completely unilluminated. We do not fundamentally advance the state of forensic science if we continue to survey examiners' mere opinions without addressing the crucial questions of how one is actually to go about evaluating observations made during comparison work (e.g., which features to select, how to assess their discriminative capacity, etc.). Similarly, few would argue that in order to advance the understanding of how to add 2 and 2 we should collect and summarise the responses of individuals who merely answer the question intuitively. Clearly: "It would be better to teach them arithmetic" [69, at p. ix]. As an aside, note that this critique of descriptivism also applies to black box studies, a type of ground truth testing prominently brought (back) to the attention of the forensic community by the PCAST Report [77]. Although of value for providing an instant view of the average performance of a given examiner or a group of examiners in a particular area of forensic expertise – which may represent useful information for assessing admissibility – traditional black box studies only record the examiners' outputs (framed in "identification language") and the congruence of these outputs with respect to ground truth. This is of rather limited scientific value. What is more, using a reporting format that is anchored in disputable "identification language" renders such studies unscientific by design.

The criticisms above should not be understood, however, as a rejection of descriptivism in principle. Some studies look beyond treating human respondents as automata that render outputs on a predefined scale. They do so by using open-response questions regarding examiners' reporting frameworks and language (which may be of any kind), and examiners' understanding thereof (see, e.g., Ref. [96]). Such studies provide a diagnosis of where a given field of practice currently stands, and where improvements should be made. This is in stark contrast to black box studies that are predicated on accepting the status quo of reporting in identification language as the relevant reference point.

It is acknowledged here that descriptivism is hardly avoidable: after all, the present paper, too, involves descriptive elements. It is also acknowledged here that the critiques of descriptivism focus on aspects that go beyond the aims and scope of the various papers that have been cited above. Notwithstanding, the main concern here is that choosing a non-scientific conclusion scale as an object of empirical study contributes to subtly perpetuating the idea that the use of such a conclusion scale is acceptable and that it could lead to scientifically warranted conclusions. The point of this paper is that treating or misunderstanding the *is* as the *ought* undermines the "science" in forensic science which, in turn, compromises the scientification of forensic practice.

3. Diagnosticism

3.1. *Defining diagnosticism in the context of forensic identification: overview and critique*

In the previous section, it has been argued that one way in which current forensic science literature contributes to perpetuating the practice of reporting identification conclusions is by considering them to be part of a valid professional toolkit. Indeed, many forensic examiners use identification language in their day-to-day reporting practice. In essence, such publications report on subjecting examiners to test cases for which they render conclusions in terms of (varying degrees of) identification/exclusion, and then summarise the proportion of conclusions in each reporting category for known same- and different-source pairs, respectively. This amounts to an essentially descriptive account of forensic identification.

Another strain of published research, called here *diagnosticism*, uses descriptivism as a starting point and seeks to *legitimise* reporting language in terms of identification within a framework of classic diagnostic reasoning. More specifically, this strain of research amounts to working out what may be called the "diagnosticity" of forensic examiners' asserted identification conclusions. For example, when a forensic examiner reports "identification", the analytical framework of diagnostic reasoning could help answering the question – so the idea – of how probable it is that a given person or object of interest is the source of a given trace, mark or impression. Research predicated on this kind of diagnosticism represents our second example for a way in which current forensic science literature keeps unwarranted identification claims in

[12] http://abfo.org/wp-content/uploads/2012/08/ABFO-Standards-Guidelines-for-Evaluating-Bitemarks-Feb-2018.pdf (last accessed January 22nd, 2022).

[13] As an aside, note that value of evidence statements are more readily amenable to logical combination, while direct opinions about propositions (i.e., source conclusions) are not [97].

existence. As will be argued below, attempts to enact diagnosticism in forensic science can run into deep conceptual problems and a series of side-effects that are best illustrated by example. Here, the recent paper by Smith and Neal [90] will be used to explain concerns about diagnosticism. For another example of diagnosticism, but with less technical details than [90], see Ref. [40].

Smith and Neal focus on "forensic science 'matching' techniques", that is "forensic procedures that involve making 'match' decisions between a crime-scene sample and a sample from the suspect" [90, at p. 319].[14] The authors discuss two notions, discriminability and reliability. They define the former, discriminability, as the capacity of a technique or procedure to distinguish between what they call "matches" and "non-matches", i.e. the event of the compared materials coming from the same and different source, respectively. By reliability the authors mean the probability of the event that compared materials come from the same source ("match") *given* that the examiner has reported "match" (i.e., the conclusion called "identification"). The authors invoke the diagnostic concepts of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), along with notions from signal detection theory.

On a broad view, the conceptualisation by Smith and Neal [90] is not fundamentally new. Yet, the authors state their account in a slightly distinctive way that leads to a several problematic assertions. The sections below outline and discuss these problems.

3.1.1. Definition of the starting point: distinguishing between the internal and the external view

In essence, Smith and Neal argue that a distinction should be made between, on the one hand, the probability of an examiner's response given ground truth and, on the other hand, the probability of ground truth given an examiner's response. This point has been made repeatedly in forensic and legal literature since at least three decades now [103]. The point is most commonly known as the avoidance of the transposed conditional [45]. However, when taking a closer look, one can note a peculiarity in the way in which Smith and Neal state their framework. They inquire about the probability of encountering the *examiner's response*. This amounts to an external perspective: i.e., taking the viewpoint of an external observer who looks at the examiner as a provider of output, and then asks "what is the probability of the expert providing a particular output (here: conclusion) given that the compared items come (do not come) from the same source?". In this perspective, the examiner is treated as a black box, comparable to and exchangeable with an abstract device or machine.[15]

This differs from the viewpoint adopted in more traditional forensic science literature, and leads to a critical problem. In traditional forensic literature on evaluating forensic science results, the focus is *not* on the examiner's response (e.g., identification or exclusion), but on the examiner's first-hand *observations* and *findings* – often referred to as the *evidence* or results – given competing propositions about the source of the evidential material. Let us call this the internal perspective here because, fundamentally, this view refers to how examiners view the problem of assigning probative value to *their* observations in the light of contrasting propositions. Stated otherwise, the starting point here is the examiner's observation of analytical features in the examined material, and the question is how to quantify the diagnostic capacity of the observed configuration of features (e.g., a given arrangement of features in a friction ridge mark) in the particular case at hand. The focus here is on the examiner using scientific knowledge and expertise to assign a value to an observed aspect of the examined item. Let us now look at some implications of the distinction between the internal and external perspective.

The level of scrutiny pursued by the internal perspective is absent in the account of Smith and Neal. Their *external* perspective focuses only on the aggregate-case probability of an examiner rendering a particular type of conclusion in test cases of a given kind, i.e. known same- or different-source pairs, eventually graded further in terms of degree of difficulty (e.g., close non-matches [65]). This way of looking at the problem of expert evidence amounts to treating the *event* of the expert rending a given conclusion (e.g., "identification") as the evidence, and assigning it a probability based on data from ground-truth testing. Specifically, this probability is equated with the proportion of times the expert concluded "identification" for test cases in which the compared items come from the same and different source, respectively. This leads to standard summary statistics such as sensitivity and 1-specificity. Clearly, and this cannot be stressed enough, these aggregate-case performance metrics tell one *nothing* about the informative content of the actually examined evidential material: the evidential material may be anything, from a low-quality fingermark with only a few poorly visible features to a high-quality mark with dozens of minutiae. Note that the internal perspective is different in this regard as it focuses on assessing the informative content of configurations of features (e.g., minutiae) *observed in the instant case* – i.e., the actual findings in the first place. The external perspective is blind to this.

It follows from the above that adopting the external perspective leads to an artificial characterisation of the value of *specific* expert evidence. To understand why reducing the evaluation of expert evidence to aggregate-case statistics (from black box studies) leaves the informative content of the examined trace material unaddressed, consider e.g. the case of an examiner (method or technique) that claims or is considered to have a sensitivity of 0.99 and a false positive rate (i.e., 1-specificity) of 0.001 (i.e., one in thousand). Stated otherwise, the probability for such an examiner to report "identification" is 990 times greater *if* the compared items come from the same source than if they come from different sources. But suppose now that the observed configuration of features in the examined mark is only moderately discriminative, e.g. the probability of observing the *features* in the mark *if* the mark came from a person other than the person of interest is merely one in hundred. In such a case, quantifying the value of the expert evidence in terms of the expert's average performance characteristics (i.e., sensitivity and 1-specificity) will overstate the probative value, compared to the actual – and more limited – informative content of the trace.[16] The reverse, though, may occur, too. Thus, two aspects are important: the rarity of the analytical features of the examined trace or mark, and the performance characteristics of the examiner. One way to coherently approach these two components has been proposed by Thompson et al. [104] (see also [98] for a representation in terms of a graphical probabilistic model). But even if Smith and Neal transitioned to the more general account of Thompson et al. [104], their focus on *rates* of false positives would still be insufficient for the need. In fact, what is needed in individual case assessment is not the aggregate-case proportion (i.e., rate) of false positive outcomes, but the *case-specific probability* of the expert reporting "correspondence" when in fact there is no correspondence between the compared items, i.e. the so-called "false positive probability, *FPP*" [104, at p. 54]. The deeper problem here is that a rate, or relative frequency, as emphasised by Smith and Neal, is not the same as a

[14] See Section 3.2.1 for a critical discussion of the term "match".

[15] Left aside here are situations in which the examiner's response is merely a 1:1 translation of the "naked" outcome of a diagnostic test, such as a spot test, used for detecting target substances (e.g., blood, saliva, drugs), reported factually. This type of examination is concerned with classification (i.e., assignment of an item or object to a particular class) rather than identification (individualisation). See Section 2.1 on the difference between classification and identification.

[16] Critics may argue that this drawback may be overcome by conducting black box testing under varying testing conditions (e.g., reflecting different case/trace types, levels of difficulty, etc.). But this objection is self-defeating as the number of case types is potentially unlimited. Moreover, the aggregate-case perspective, by design, does not quantify the value of the trace-related features.

probability [68], a point to which we will return later in Section 3.2.4.

Let us attend to the above points in further detail and highlight some of the consequences of our argument. Smith and Neal insist on computing the probability of ground truth (here: the proposition that the compared items come from the same source) *given* the examiner's report of an identification, using Bayes' theorem. Let us be precise about what this posterior probability, call it S/N, means. It is:

> S/N: The posterior probability of the proposition that the compared items come from the same source given *that the examiner (method or technique) has concluded "identification."*

This is an artificial, case-alien probability because it amounts to interpreting the value of (fingermark) evidence in terms of the expert's (method's or technique's) *general* performance characteristics only. Most importantly, the practical consequence of using S/N would be to draw the *same* conclusion in all cases where the expert reports "identification", disregarding, thus, the informative content of the configuration of features of the mark examined in the case at hand.[17] It is doubtful whether consumers of expert evidence are interested in such an abstract matter, let alone whether they should be encouraged to believe that a S/N posterior probability suitably reflects what they should conclude from the examined evidential trace, mark or impression.

Some quarters of forensic science have long been aware of this problem. Their research efforts have focused on methods for evaluating the capacity of analytical features in traces, marks and impressions – not merely an expert's abstract assertion of "identification" – to help discriminate between competing propositions regarding the source of evidential items. The metric they have developed for quantifying the value of forensic observations and findings is the likelihood ratio [e.g., 2, 81]. It is actually both surprising and alienating that Smith and Neal manage to conceal, in their paper, this strain of existing theory based on the likelihood ratio. The notion of discriminability that Smith and Neal discuss for expert assertions of "identification" (and degrees thereof), seemingly as a novelty, is essentially the same as what traditional forensic literature on evaluation of findings is doing for decades, though with a focus on analytical features of evidential materials. See e.g. Refs. [42,43,91], the supplementary materials of [52] for an annotated historical overview of relevant literature on the use of the likelihood ratio for evaluative purposes in forensic science, and [79] for an account of performance assessment of likelihood ratio procedures. The merit of this perspective is to clarify the idea that the proper role of the forensic scientist is to focus on the value of examined items and materials – i.e. their analytical features – and to abstain from expressing direct opinions about source propositions [109].

In summary, the above considerations imply two main drawbacks for Smith and Neal's account. Considering an expert's conclusion in terms of "identification" language as the starting point is, first, prone to conveying that the expert is expressing a direct opinion on a proposition (i.e., a ground truth state), which is beyond the expert's area of competence. Second, their account is uninformative about the actual evidential material, i.e. its analytical features. It is not suggested here, however, that an expert's *general* performance characteristics are irrelevant altogether: these metrics may be of value for characterising the qualification of an expert for a particular task *in general or on average* (as compared to laypersons), which may be a relevant consideration at some stage in legal proceedings (e.g., questions of admissibility). It is maintained here, however, that such general performance descriptors are uninformative *by design* about the evidential value of any particular trace, mark or impression. Specifically, the computation of a S/N posterior probability, as defined above, is answering an abstract question, disconnected from the actual case.

Smith and Neal concede that their account is not readily applicable in an operational environment ("We are far from being able to confidently apply this framework in the courtroom" [90, at p. 330]). This can be agreed with, but not because there is a lack of data, as Smith and Neal argue, but because their account does not focus on quantifying the value of the actual evidence. The lack of data of the kind that Smith and Neal have in mind leads them to assert that "the scientific study of forensic science is still in its infancy and there is a lot of work to be done before estimates of reliability in a given case can be made with any level of precision" [90, at p. 330]. This misrepresents the state-of- the-art of forensic evaluation methods based on the likelihood ratio, as referenced above. Examples are, just to name a few, [73] in forensic voice comparison, [101] in DNA analysis and [75] in fingermark analysis. In the particular case of probabilistic genotyping systems for analysing the results of DNA mixtures, a recent NIST report [22] found that there are at least 60 publications that contain some form of validation data.

3.1.2. The traps of attempts to conceptualise identification as an instance of classic diagnostic testing

Proponents of diagnosticism portray forensic identification (e.g., in friction ridge analysis, toolmark examination etc.) as an example of a "diagnostic testing procedure" [90, at p. 319]. Though a seemingly telling description at first sight, the analogy between the problem of forensic identification and standard diagnostic testing can be challenged. To explain this point of view, let us start by stating the main aspects of classic diagnostic testing.

According to medical literature, a diagnostic test is, broadly speaking, designed to "identify" individuals with a target condition [e.g., 30]. A diagnostic test is part of a process that seeks to classify individuals into categories. These categories designate particular (pathological) conditions. Both categories and conditions are defined and agreed upon by a relevant scientific community. In the simplest case, a diagnostic test and results of related diagnostic accuracy studies can be thought through in terms of a 2×2 table (a four-cell matrix): there are two columns representing, respectively, the presence and absence of the target condition in the examined individual, and two rows representing the test outcomes, commonly referred to as "positive" and "negative" respectively. It should be emphasised that this is a simple description because, in practice, diagnostic tests may not always render clearly positive or negative outcomes [e.g., 87–89]. The binary simplification is, however, no detriment to the generality of the argument pursued here. The question to investigate now is whether this classic diagnostic setup suitably captures the problem of forensic identification.

The classic diagnostic setup is readily illustrated in applications where the purpose is to recognise individuals or items in a population of individuals or items that belong to a certain category (or, class). All members of the category of interest have a given target condition or property. For example, one may wish to "determine" whether a given fluid is blood (or sperm, saliva, etc.), or – in a medical context – whether a person is infected by a certain type of virus. To "detect" the target condition, a diagnostic test will focus on one or more features thought to be systematically associated with the members of the class of interest. For example, tests for human pregnancy commonly focus on detecting human chorionic gonadotropin (hCG). While different tests may vary in their performance of "detecting" target features, by providing some sort of signal, the important point for us here is a more general one. The point is that there is a well-defined feature (or combination of features), the detection of which is used to ascribe persons or items to a particular class. Designing diagnostic accuracy studies for this type of problem is readily feasible: it suffices to arrange test items with known group membership status, subject them to the testing procedure and record the test results (e.g., presence or absence of signal). Conducting such studies

[17] Smith and Neal might object to this by arguing that they are not only considering the conclusions "identification" and "exclusion", but also intermediate labels reflecting "different degrees of 'matchingness.'" [90, at p. 324]. Degrees of similarity are quantified elsewhere in forensic literature in terms of similarity metrics (also called scores), but this is not what Smith and Neal deal with. They focus on discrete labels, assigned by examiners, as surrogates of degrees of similarity.

is also fairly standard and smooth in the sense that all the testing process essentially needs to do is concentrate on the *same pre-defined class characteristic(s)*. Coarsely speaking, the aim is to design the test in such a way that it yields a signal whenever the class charactistic(s) is (are) present (and discernible/detectable), keeping an eye on metrics such false positives and false negatives.

This paper contends that the above diagnostic testing framework has little to do with forensic identification (or, forensic inference of source) and that a considerable bend needs to be operated to claim that forensic identification methods are, as Smith and Neal take as a premise, an instance of diagnostic testing. For a deeper understanding of this point, consider again the generic example of an examiner who compares a fingermark found on a crime scene with a reference print from a person of interest. The examiner observes a certain number of similarities and differences between the mark and the reference print. It is unknown whether the mark comes from the person of interest, or from an unknown person. To conceptualise this as an instance of diagnostic testing, one needs to cope with two aspects: one is framing the target class, the other is defining the feature(s) thought to be "indicative" of that class. Both of these aspects, if bent to the above account of classic diagnostic testing, lead to considerable mind contortions that, as argued below, epitomise the problems outlined in Section 3.1.1.

Start with the first aspect: is there something like a "class" in forensic identification? Let us assume that proponents of diagnosticism might say that the target class is the population of mark-print pairs, for each of which both mark and print come from the same source. That is, querying whether the mark-print comparison at hand is an instance of the population of same-source pairs. This would emulate the focus of a diagnostic test, which is to recognise a person or item as a member of a particular class (e.g., classifying a biological fluid as blood). Thus, the class feature in forensic identification would be the fact that the compared items come form the same source. While this sounds overly cumbersome, let us *provisionally* accept this idea of "same-sourceness" as a category property and move on to the second aspect. It requires us to ask: is there a defining feature of "same-sourceness", akin to the human pregnancy hormone (hCG), on which the diagnostic testing procedure could focus? It is here that the idea of conceptualising forensic identification as a problem of classic diagnostic testing collapses. In forensic identification, there is *no* class-wide, standard analytical feature that all same-source pairs have in common. And the reason for this is that, strictly speaking, each same-source pair *defines its own category*. Forensic identification – i.e., individualisation – is classification for categories defined by a single item. Theoretically, each same-source pair has, in its entirety, its own feature set, *to the exclusion of all others*.[18] This renders forensic identification much more challenging than what the idea of classic diagnostic testing suggests. While in classic diagnostic testing it is sufficient to "detect" an agreed-upon class-wide feature that *many* people (items) may possess, forensic identification encounters the fundamental problem of determining whether a given detected combination of analytical features is *uniquely* indicative of a *single* source. Thus, in all cases where the pool of candidate sources cannot exhaustively be investigated, forensic identification must rest upon an unprovable claim of discernible uniqueness.

The problems of the supposed analogy between classic diagnostic testing and forensic identification run even deeper. Recall that in conventional diagnostic testing, such as pregnancy testing or body fluid "classification", the target analyte for each category is *pre-defined exactly* (e.g., hCG). But in forensic identification, there is no analogue to this. One simply cannot tell, for any candidate source (e.g., a finger, screwdriver, etc.), what the feature set in a trace or mark left by the candidate source will *exactly* look like.[19] Depending on factors such as the angle and pressure, and subsequent external constraints (e.g., exposure to environmental conditions), there will be an inevitable variation in the combination of features present in *different* marks, traces or impressions *even though* they come from the same source. Thus, if one cannot even tell, a priori, what the feature combination in a trace of a any given source looks like, it remains unclear how forensic examiners can claim, when they observe particular features (e.g., striations in a toolmark, minutiae in a fingermark), that those features are *uniquely* "pointing" towards a particular candidate source.

However, diagnosticism, as Smith and Neal frame it, stays away from the above problem of definability, recognisability and quantifiability of features for classes thought to contain a single member. Instead, diagnosticism treats all identifications in the same way, by taking the expert's assertion of "identification" – akin to a "positive" diagnostic testing result or the wag of the tail of a (drug) detection dog[20] – as the sole indicator of "same-sourceness". Thus, by reducing scientific evidence to an expert's mere statement, diagnosticism amounts to turning a blind eye to the actual trace material and the probative value of its features.[21] This brings us back to the main argument of this paper: publications predicated on this sort of diagnosticism do not help advance the fundamental understanding of the evidential value of forensic traces, impressions and marks. It keeps the conceptual framework limited to nothing better than the level of scrutiny one can apply, e.g., to detection dogs. This massively undervalues human intelligence. Most importantly, diagnosticism treats forensic identification in terms of *aggregate-case expert performance* characteristics whereas, in reality, the fundamental problem of inference of source is the empirically unsurmountable justificatory burden deriving from the notion of "to the exclusion of all others".

3.2. Further side-effects of misconstrued diagnosticism

Diagnosticism in literature on forensic identification is typically accompanied by a series of topics pertaining, among others, to terminology and the concept of probability. When dealt with in a confusing way, these topics make diagnosticism even more convoluted and more laborious to deconstruct. The sections below exemplify a few of these topics, using again Smith and Neal's [90] paper as an example.

3.2.1. The problematic nature of the "match" paradigm

The terms "match" and "matching" – as an adjective, a noun and a verb – are so commonly used in forensic science literature and practice that the reader may wonder what could be wrong with these terms. The short answer is: almost everything, which is why these terms should have no place in forensic science,[22] despite the fact that they are prominently used throughout official documents, such as the PCAST Report [77]. Longer answers have previously been given by other authors [47,72], making it all the more exasperating that "match"-terminology continues to be highly prevalent in current forensic science literature.

[18] Note, however, that in practice, the features of a source do not necessarily appear faithfully (i.e., accurately) and completely in *traces*. Moreover, examiners have less than perfect capacities to properly discern feature sets, thus compromising the possibility of assertions of the kind 'to the exclusion of all others'.

[19] This includes DNA traces. Even though there are standard sets of markers (loci), the actual allelic configuration of an individual source at each of the selected markers is, a priori, unknown. In addition, perturbing phenomena such as drop-in and drop-out may occur [e.g., 21].

[20] This analogy has previously been made in a presentation by Christophe Champod held at the National Commission on Forensic Science Meeting #12, Washington, DC, January 9, 2017.

[21] Research on forensic interpretation currently addresses this topic using so-called feature- and score-based likelihood ratio procedures (see also discussion and further references in Section 3.1.1).

[22] See also Thompson [102] ("forensic scientists should use this term [match] cautiously, if at all, when reporting their conclusions" [at p. 797]).

A main problem of the term "match" is that it is used to denote two fundamentally different targets: observations on the one hand, and ground truth on the other hand. This makes it difficult to know, at any one point in time and without additional explanation, what exactly discussants mean. Consider this in the context of Smith and Neal's account of forensic diagnosticism. In the introduction of their paper, they state: "The present work focuses exclusively on forensic procedures that involve making 'match' decisions between a crime-scene sample and a sample from the suspect." [90, at p. 319] This sentence suggests that the term "match" is a descriptor of the extent of agreement between the observable features of the two compared items, assessed and declared by the examiner. As Evett et al. [47] put it: "The match paradigm calls for a judgement, by the scientist, as to whether or not the two sets of observations agree within the range of what would be expected (…)" [at p. 18]. But this is not the only way in which Smith and Neal use the term. They also assert that "science ought to be focused on measuring the extent to which a procedure can discriminate between two classes, 'matches' and 'non-matches' (discriminability)." [90, at p. 319] In this sentence, the term "match" refers to the proposition that two compared items (objects, traces, marks or impressions) come from the same source, i.e. a ground truth state. The amalgam of the two meanings is most visible when Smith and Neal conceptualise forensic identification as a "rating task", i.e. a "2 (ground truth: match, non-match) x 2 (decision: match, non-match) confusion matrix" [90, at p. 324], and when they discuss the computation of posterior probabilities $p(M|D)$, where M denotes "match" and "D" the scientist's assertion ("decision") that there is a match. This leaves the lay consumer of expert evidence wonder why, on the one hand, an expert asserts "match" (understood as observation) while, on the other hand, the expert's assertion does *not* mean that a match (understood as ground truth) is certain, but at best probable to some extent.

The discrepancy mentioned above is all too well known in connection with the similarly problematic term "identification". For example, the DOJ's ULTR, defines a "source identification" as "an examiner's conclusion that two friction ridge skin impressions originated from the same source" [106, at p. 2] but, at the same time, insists that "an examiner shall not (…) assert that two friction ridge skin impressions originated from the same source to the exclusion of all other sources" [106, at p. 3]. It is hardly surprising that the reactions to such statements have been incisive and sharp. For example, during a meeting of the Advisory Committee on Evidence Rules, it has rightly been "queried how an examiner logically could state that a mark came from a particular defendant without saying it didn't come from another person" [1, at p. 20]. It has even been suggested that the language amounts to "rhetorical chicanery" [1, at p. 21]. Indeed, as mentioned previously in Section 2.1, in most practical cases examiners cannot "exclude all others", hence the problem is the use of language in the first place that suggests the contrary. This is as much the case for the term "match" as it is for the term "identification".

On a more conceptual side, two additional complications are worth mentioning. First, the term "match" suggests something like "identity" or "being identical", or even "identification",[23] yet in forensic comparison work this is an impossibility. Because, by definition, an object can only be identical with itself, two items cannot be identical with one another *even* if they come from the same source.[24] Second, as argued by Morrison et al. [72], the term "match" is particularly unsuitable in cases where the data relating to the measurement of features are not discrete, but continuously-valued. Interestingly, the common understanding of the term "match" actually admits that the nature of data is not discrete, otherwise the "match paradigm" would not require, as noted in the above quote from Evett et al. [47], a judgment by the scientist.

Sceptics might invoke that the notion of "match" is not limited to a binary understanding in terms of "match/non-match", but includes, using Smith and Neal's words, the idea of (degrees of) "matchingness" [90, at p. 324]. This, however, is only a lip service because Smith and Neal suggest to implement this notion, alas, through, a *discrete* scale of categories of "matchingness". It is not clear how this ought to be implemented because if – as contended here – there is no unique way of defining when a degree of similarity between compared items constitutes a "match" rather than a "non-match", this difficulty will only be exacerbated when considerations are extended from binary- to (discrete) multiple-category scales. In this sense, the "match" paradigm cannot serve as a substitute for attending to understanding the nature and probative value of analytical features in the first place, which is a necessary requirement for the explainability of scientific evidence.

In summary, thus, one can see that there is no dimension in which "match"-terminology proves useful. It is internally inconsistent and not conducive to the attainment of the aims it purports to achieve. Most concerning is that "match"- terminology is one way in which diagnosticism is given a seemingly formal framework whereby unwarranted claims of forensic identification are being perpetuated in forensic science literature.

3.2.2. Forensic "prediction" techniques?

A further side-effect of both diagnosticism and its manifestation through "match"-terminology is the confusion that reigns around the term "prediction". Smith and Neal, for example, assert that "forensic science 'matching' techniques" [90, at p. 319] are "used to *predict* ground truth" [90, at p. 319; emphasis added]. While the appearance of the term "prediction" is little surprising in a context marked by diagnosticism, with its core notions of positive/negative *predictive* value, this is not an excuse for incorrectly using the term in forensic science applications [8,15]. The reason for the unsuitability of the term "prediction" as a descriptor of a forensic examiner's conclusion in the context of inference of source should be obvious: the issue of whether or not a given trace or recovered material comes from a candidate source is a *fixed* matter of the present, which is the exclusive consequence of an event that happened in the past. A source proposition in forensic inference bears no relationship with a hitherto unrealised event to which the term "prediction" could allude to. For the same reasons, it should be obvious that there is no place for the term "prediction" in matters such as inference about the cause of death [53], the position of individuals in a car prior to a road accident [18], or externally visible features of donors of DNA [61,107], just to name a few examples from forensic science literature that demonstrate the widespread confusion about this term.

3.2.3. Unfounded carrying-over of source conclusions to ultimate issues

The reader might find this paper's insistence on the (logical) impossibility and unsuitability of forensic examiners' opining on source propositions (Section 2.1) and the use of 'match'-terminology pedantic and, after all, a relatively minor problem. This, however, would misconceive the seriousness of the matter because the truth is that the problems run much deeper: in particular, questions of source are prone to be carried over to ultimate issues, such as (criminal) liability. While evidence for this concern among laypersons is more anecdotal than based on hard evidence, its is known that academics can fall for this confusion, thus giving us reason to fear that it might also occur among laypersons.

Consider the following assertion by Smith and Neal: "through use of Bayes' Theorem, the forensic scientist can shed light on the answer to the question that the criminal justice system wants to know: what is the probability that the suspect is guilty?" [90, at p. 321] Just to make it clear for the readership, the suggestion here is, literally, that forensic

[23] An example for equating the term "match" with "identification" (in the sense of same source) is the statement: "The examiner (…) determines whether the latent print from the crime scene matches the source print provided by the suspect (*identification*) or not (*exclusion*)." [90, at p. 320; emphasis as in original].

[24] On this point, see also the argument presented in Section 3.1.1 on the a priori undefinability of the feature configuration in traces.

scientists compute a probability of "guilt". Thus, the paper provides an example for the subtle shift from considerations of source (or, "match" in Smith and Neal's terminology) to the ultimate issue (guilt). Their paper starts by explaining the framework of diagnostic testing using propositions of source, including notions such as positive/negative predictive value, and then – in later parts – refers to the same propositions in terms of "guilt". These ideas are problematic in at least two ways, not to mention the fact that opining on ultimate issues by scientists is procedurally barred.[25]

First, it is important to recall and understand that aggregate-case metrics of diagnostic performance, derived from controlled experiments using known same- and different-source pairs as assumed under Smith and Neal's account, cannot – by definition – *directly* serve as quantifiers of probative value with respect to ultimate issues. This does not mean that inference of source cannot be extended to inference about ultimate issues. This is possible, theoretically, but requires a more elaborate probabilistic analysis, already reported in forensic science literature in the early 1990s [44,92,94]. What these developments show is that extending considerations from source to what is also known as "crime level" [33] requires more than information regarding the relative rarity of the analytical features (of the trace material).[26] In essence, there are two additional factors to consider. One concerns the question of whether the recovered trace material is relevant, where "relevance" refers to the question of whether recovered material comes from the offender, i.e. thus helping in the consideration of persons of interest as possible offenders.[27] The other factor refers to the event of material coming from the person of interest even though the material is not related to the event under investigation (i.e., the material is not relevant in the sense defined above) and the person of interest is not the offender. For a reconstruction of these probabilistic developments, using graphical probability models, see e.g. Ref. [50]. Clearly, these considerations are above and beyond aspects characterising the expert's comparison task, hence aggregate-case metrics of diagnostic performance of the kind advocated by Smith and Neal come not even close by the requirements for logically extending inference of source to higher propositional levels (such as "crime level"). Hence, one can but conclude that Smith and Neal's assertion that error rates considered in terms of positive/negative predictive value from classic diagnostic testing (i.e., predicated on varying assumptions of source) "would directly inform on how strongly the evidence implies that the suspect is guilty or innocent" [[90], at p. 329] is simply incorrect.

Second, Smith and Neal's suggestion that Bayes' theorem may be used to compute "the probability that the suspect is guilty" [90, at p. 321] is technically, conceptually and definitionally improper because "guilt" is not a proposition, but – as repeatedly pointed out by legal scholars [e.g., 3,4] – a legal conclusion, i.e. a *decision*.[28] While there is a way to approach *decisions* regarding ultimate issues in a formal analysis, this requires more than probability theory; it requires decision theory [60,63], which can also be applied with a more limited focus to treating (forensic) source conclusions as decisions [12,16,32]. Further discussion of this topic is presented later in Section 3.2.5.

3.2.4. Probability conundrums

The suggestion of Smith and Neal that scientific evidence could lead to a posterior probability of "guilt" is not only problematic in the way explained in the previous section. Smith and Neal take the idea of a posterior probability a step further and conjecture about complementing probability with an interval. More specifically, they write: "Eventually, the field might be able to provide a confidence interval of sorts regarding the probability for a given case (…)" [90, at p. 330]. This proposal is objectionable on two grounds and should be advised against. First, trying to fuse the core Bayesian notion of posterior probability with the frequentist concept of confidence interval is a contradiction in terms (see Ref. [62] for a detailed discussion). Second, a probability for a single non-repeatable event is not an interval, but a *single* number, expressing one's uncertainty about the truth or otherwise of this event; *different* numbers (probabilities) – by definition – express *different* states of uncertainty. One may not find it easy or be unwilling to pin down one's probability in terms of a single number, but this does not imply or suggest that there should be a "confidence interval of sorts" [90, at p. 330]. Likewise, one should also resist the temptation of placing a probability on a probability as the loose idea of "confidence in a probability (assertion)", widely used in informal discussion, might suggest.[29] In view of this, scientists should carefully choose their reporting language. For example, they may use expressions of orders of magnitude,[30] but they should not use a *perceived* difficulty in applying the concept of probability as a reason to tweak this concept in incoherent ways (e.g., adding a confidence interval), so as to bend it towards mere intuition. Probability is a normative, not a descriptive framework [14].

There is yet a further problem in the way in which Smith and Neal conceive of the "probability of guilt". They spend considerable effort discussing the role of base rates in the computation of positive/negative predictive value and $p(M|D)$, suggesting that the base rate *is* or serves as a (suitable) proxy for the prior probability. Similar ideas are widely advocated in legal literature [e.g., 34], but present a series of shortcomings [e.g., 11]. Exposing these shortcomings in detail is beyond the scope of this paper and redundant with respect to existing literature on this topic. We shall only recall one strain of argument from a definitional point of view. Consider that the notion of base rate refers, broadly speaking, to the proportion of a population that has a given feature (or condition). This is readily understood in the context of medical diagnosis where the focus of inquiry is the prevalence of a certain disease in a population of interest. One can also think this notion through in terms of inspecting members of the target population, leading to frequency data. Likewise, a forensic scientist might inquire about the proportion of a population that has a certain blood type. However, the *relative frequency* of a feature in a *sample* from a target population is conceptually different from the *probability* that a given member of the target population has the feature of interest.[31] Going from the former to the latter requires additional argument and assumptions [99]. What is more, in the legal context, conceiving of probability regarding the ultimate issue in the individual case, using frequentist ideas, has repeatedly been exposed as unworkable [e.g., 6,49,60,67]. Of course, one is free to ignore these challenges, but it is clearly insufficient then to leave readers alone with assertions such as:

> "(…) the base rate in the real world is unknown (…). We simply do not know how often the police suspect actually is the culprit.

[25] For a discussion of and further references on the ultimate issue rule, see e.g. Robertson et al. [82, p. 50–54] and Dennis [38, para. 20–022].

[26] Recall that, as noted in Section 3.1.1, Smith and Neal's account does not focus on quantifying the diagnostic capacity of analytical features, but only on aggregate-case performance metrics.

[27] For a development in cases where "source" refers to an object (e.g., shoe) rather than a person, see Evett et al. [46].

[28] As an aside, this also renders the idea of a "base rate" [90, at p. 326] (of guilt), to which Smith and Neal's expression "how often the police suspect actually is the culprit" [90, at p. 326] alludes to, vacuous.

[29] On this point, see also Lindley [68]: "(…) it is nonsense for you to have a belief about your belief if only because to do so leads to an infinite regress of beliefs about beliefs about beliefs" [at p. 115]. For a discussion of this point in the legal context, see Ref. [11].

[30] On the notion of orders of magnitude of probabilities and likelihood ratios see, for example, the ENFSI Guideline for Evaluative Reporting in Forensic Science [109].

[31] For further discussion, see also de Finetti [36, at p. 128]: "The frequency with which certain events obtained or will obtain cannot be identified with probability. Frequency is a mere fact, independent of both the meaning of probability and the probability values assigned to the events."

> Exacerbating this problem, there is no single base rate; the base rate varies across jurisdictions, police departments, divisions within the same police department, and even across individual officers." [90, at p. 326]

This is tantamount to introducing a concept and, at the same time, admitting that it is not operational.[32] It is worth noting that the conceptual limitation of the chosen framework is an instance of the deeper and unsurmountable problem that the frequentist perspective does not provide an operational definition of probability [66]. Smith and Neal's suggestion that readers could explore "posterior by prior curves" [90, at p. 326], i.e. considering a range of prior probabilities and displaying the corresponding posterior probabilities, offers no help. It merely exemplifies Bayes' theorem, which is uncontested, and hence a point that does not need to be made. Insofar, posterior by prior curves amount to treating the symptoms rather than addressing the root causes. The unresolved problem remains how to choose a prior probability in the first place. Interestingly, Smith and Neal ask "how are we to *decide* which base rate to factor into our calculations (…)?" [90, at p. 326; emphasis added], suggesting that probability assignment is a decision. While there is, actually, an important theoretical account that considers probability assertion as a decision [e.g., 35–37],[33] Smith and Neal do not go along that route. Little surprisingly, the decisional account of probability assertion would direct us to considerations quite different from base rates. It requires one to admit that probability assignment is, first of all, a personal decision for which one needs to take responsibility. *Deciding* to assign a particular probability, in this perspective, is informed by data (see Ref. [99] for a discussion), but does not reduce to data *only*[34] as suggested by the (simplistic) equation of probability with base rate.

3.2.5. *The vague notion of "decision criterion"*

Diagnosticism typically refers to examiners' conclusions as decisions. For example, as noted earlier, Smith and Neal discuss "forensic procedures that involve making 'match' decisions" [90, at p. 319]. This is in line with mainstream parlance that uses "decision" as a fashionable term, developed in forensic literature over the past decade, but revealed as a way to circumvent the justificatory burden associated with the idea of identification in the sense of "to the exclusion of all others" (see, in particular [31], and, more recently, [32]). That is, rather than providing a rationale for a proffered conclusion (e.g., "identification"), reference is made to a conclusion as a decision, adding that examiners have been trained and shown to be able to make such decisions reliably – exactly as is stipulated by diagnosticism.

Let us now address a further problem of discussions that refer to forensic identification as a decision: the use of the vague notion of "decision criterion". This notion is vague because it is often used as though it were clear what this term meant. Consider, for example, Smith and Neal's assertion: "The examiner's decision criterion is the amount of information the examiner requires to make a particular classification decision." [90, at p. 322] While, at first sight, this sounds sensible, this can be seen as merely emulating an intuition drawn from day-to-day decision making; i.e., the common saying that one decides, or that one's decision criterion is met, whenever one has "enough information". This may sound elaborate, but is devoid of any substance: neither "information" and its measurement (required to give meaning to the term "amount") are defined, nor what the requisite criterion for decision *precisely* is. Many forensic science disciplines have a long history of operating upon this idea of thinking that there is a fixed relationship between (a given amount of) information and a particular conclusion, most notably friction ridge analysis with its now widely abandoned minimum number of minutiae identification standards (e.g., the so-called 12 point rule). In these accounts, the type of conclusion to be given is a rigid function of nothing else but the kind of observation made.[35] These traditional attempts to define conclusion schemes akin to a rule-based system with *if–then* clauses do not withstand scrutiny on at least two points. First, information (or, in a forensic context, an observation) is only a starting point for inference, i.e. the reasonable reasoning under uncertainty. But this remains far from a decision because inference is only a preliminary to decision. Second, *making* a decision *logically* requires one to draw one's attention to the potential decision consequences, prior to making a decision [57]. Clearly, not all potential consequences of one's actions are equally desirable, hence one's decision making framework should account for the relative (un-) desirability of decision consequences as well as for the probability with which each of those consequences are thought to occur. Hence, on pain of falling short of reality, any decision criterion for forensic identification must specify how to aggregate these fundamental decision ingredients. The prime candidate theory to do this is (Bayesian) decision theory, and it has been studied to see what kind of light it could shed on the question of decision criteria for forensic identification [12,16,100]. The conclusion is rather disappointing for proponents of "identificationism" and diagnosticism. In essence, decision theory shows us that a decision involves two elements. One of these elements is what we believe, i.e. how strongly we believe that one potential state of the world, rather than another, is true. The other element is our preferences among decision consequences, expressed e.g. in terms of utilities or losses. However, since scientists are not in a position to specify any of these ingredients,[36] they are not in a position to make any decisions.

The bottom line of this "is that experts should abandon the identification/individualisation conclusion altogether" [28, at p. 96]. Clearly, this would dissolve the diagnosticists' primary object of study – identification conclusions/decisions – which may explain why current literature largely shies away from addressing the fundamental challenges that decision theory poses to forensic identification. Instead, what one sees, are studies that proceed axiomatically as though identification, as practised by scientists, is a well-founded and admissible reporting category.

4. Machinism

4.1. *Machinism in the context of forensic identification*

The discussion of forensic identification so far in this paper is predicated on the view that identification is, in essence, a *human* activity and, thus, *imperfect* by design. This imperfection is the reason for the empirical study of the performance of human examiners as well as the development of conceptual frameworks for processing partially reliable information provided by examiners. Given the inherent deficiency of

[32] As an aside, inquiring about a "naked" base rate is pointless insofar as when scientific evidence is heard, most of the time, other evidence has already been heard, hence the starting point is not in the void. In addition, inquiries into categorising cases into abstract classes took a controversial turn in legal literature, in particular in discussions of the so-called reference class problem [e.g., 5].

[33] See also [11] for a discussion in the legal context, and [13,17] for forensic science applications.

[34] For further discussion of data-centrism, see also Section 4.3.

[35] Likewise, examiners in trace evidence disciplines have been referring to different "levels of association" [108, at p. 207] as a function of the observations made. For a critical discussion of a similar proposal in the context of shoemark analysis, see Ref. [27].

[36] Scientists are neither in a position to assert a probability for the proposition that a person of interest, rather than an unknown person, is the source of a given stain or mark, nor are they in a position to articulate utilities or losses for decision consequences (e.g., the consequence of reporting "identification" when in fact the person of interest is not the source of the fingermark). To make the latter aspect clear: since the consequences of a scientist's report, by definition, will affect a third party (i.e., a given person of interest under investigation or at trial), the scientist *cannot* be competent to evaluate the relative (un-)desirability of said consequences [10].

human expertise, it is hardly surprising to see that, in recent years, scientists brought up the idea of assigning forensic comparison work to machines, either partially or completely, including the conclusion stage. More specifically, the idea – called here "machinism" – is to design machines that can "learn" the task of forensic comparison and identification so as to eliminate the foibles of human expertise [23]. This direction of research seeks to draw advantage of the vast field of machine learning (ML), a sub-field of artificial intelligence, which currently attracts strong interest in virtually all areas of science, in particular where large amounts of data are available.

This section provides a sketch of the standard ML setting, exposes the key assumptions underlying standard ML procedures and explains why they fall short of the nature of forensic identification. It will be argued, thus, that attempts to approach forensic identification in its entirety through ML represent a further instance of the persistence of identification claims in forensic science literature.

4.2. A sketch of the basic ML setting and its position within AI

Broadly speaking, ML can be seen as a form of computer programming, intended to give computers the ability to perform certain tasks or, as some may say, cognitive abilities. ML is a sub-field of artificial intelligence (AI) [e.g., 71] which, in turn, is a discipline within computer science. It is important to understand that there are different "philosophies" within AI and computer programming. In the era spanning approximately from the 1950s to the late 1980s, the predominant (or, classic) view was that a computer program contains all the information necessary to transform inputs to outputs. Stated otherwise, it is assumed that the human programmer is effectively able to define the (cognitive) tasks the computer ought to perform. A term commonly encountered in this context is symbolic AI,[37] referring to programming instructions involving formal symbolic representations. While this perspective is well suited for certain tasks, such as logic and probabilistic reasoning, there are many other tasks that humans cannot easily define in terms of formal rules. Examples are perceptual and motor tasks. Such tasks have been approached by what is called subsymbolic AI. This approach is based on the idea of learning from experience, rather than strict symbolic representation of rules and properties. That is, the target task is thought to require knowledge that cannot directly be provided by the human programmer, but is to be extracted from input data examples during a training stage. This brings us to the topic of ML, a subfield of AI.

Within the limited scope of this paper, it is not possible to address ML in its entirety. Deep learning, a special subfield of ML, will be left aside.[38] Also not addressed here is unsupervised learning, typically concerned with tasks such as clustering (i.e., given input data-points, assigning each of the data-points to a group). Instead, we will look in more detail at *supervised* learning.[39] This type of learning is concerned with the general problem of processing inputs (features) to outputs (labels) based on examples of inputs for which the associated output labels are known (i.e., the training data).[40] Two main applications of supervised learning are regression and classification. Only the latter will be addressed here. Both, regression and classification use one or more input variables, but the two types of learning differ in their outputs. In regression, the output is real-valued whereas in classification – binary or multi-class – it is categorical. Let us now consider the general ML setting for classification.

In a nutshell, a basic ML setting consists of two functions.[41] The first is a parameterised function or *model* that maps inputs to outputs. The parameters of this function are 'learned' through training data using the function's corresponding learning algorithm. There is a difference, thus, between the function that performs classification based on some parameters (i.e., the *learned* or *trained model*), and the corresponding *learning algorithm* [e.g., 7]. As an aside, note, however, that there may also be parameters that cannot be learned from data. These are so-called hyperparameters and finding values for these parameters is called 'tuning' a model.[42] The second function of the ML setting measures the performance (or error) of the classifier by comparing the outputs (i.e., labels or category assignments) to the actual labels (i.e., *known* category membership) of the training data. The general problem is to learn the parameters of the classification function so as to optimise the chosen measure of success for the given input data.

4.3. A critique of ML as applied to forensic identification (individualisation)

At this point the reader might be tempted to think that our presentation got lost in the details and yet remained too general and incomplete on the level of specific ML models, and the relevance of these topics for forensic identification. Let us consider, thus, a commonly used classification method as an example: random forests. A random forest (RF) is a tree-based method that involves, at the training stage, the construction of multiple decision trees (forming the 'forest'). Broadly speaking, a decision tree[43] is a method for assigning an item to a class (or category) based on an item's features, by asking oneself through a series of questions. To construct an individual tree of a RF, the training data is bootstrapped. That is, only a part of the training data is selected (though a datapoint can be chosen more than once). A tree is then constructed based on a random selection of features (i.e., variables). Repeating this procedure many times creates the RF. To use the RF, a new input item is processed through all the trees in the forest and the individual classification output of each tree is recorded. The new input item is assigned to the class (category) which has received the most votes. To assess the performance of a RF, one can process part of the data that has been set aside for testing (the so-called 'out-of-bag sample'). The testing leads to data regarding the proportion of correct and incorrect classifications. See e.g. Ref. [54] for an example of the use of the RF classifier in the context of land-to-land mark comparisons on fired bullets, and [24] for source inference of ammonium nitrate.

Let us take a closer look now at the idea of applying the standard ML setting (Section 4.2) to the problem of forensic identification. Applying the standard ML scheme to forensic identification *would* mean, in the first step, selecting relevant data for the problem at hand: here, this would be – for example – data pertaining to pairs of items *known* to come from the same source, and data pertaining to pairs of items *known* to come from different sources. Often, it may not be possible to process data directly by ML models and, thus, it may be necessary to pre-process data and/or conduct feature engineering. Next, a selected ML model is trained with part of the data. That is, the model is 'fed' with (many) examples of inputs for which the category label (here: same or different source) is known. Once one has a trained model, its performance is

[37] Symbolic AI is also sometimes referred to as "Good Old-fashioned AI" (GOFAI) [e.g., 19].

[38] Note, however, that the critiques of ML developed hereafter also apply to deep learning, even to a larger extent.

[39] Other categories of AI, such as reinforcement learning, that do not fit easily either into either supervised or unsupervised learning [e.g., 74] are also not considered here.

[40] The term 'supervised' comes from the fact that training data consist of known pairs of input and output values through which a program can be 'supervised' during learning. See e.g. Ref. [86] for other dimensions in which learning types can be classified.

[41] For more technical accounts, see e.g. Ref. [86].

[42] Note that 'tuning' often is essentially proceeding by trial and error or, more colloquially expressed, 'turning the knobs'.

[43] Decision trees for classification (typically having a top-down structure) as discussed here must not be confused with the horizontally constructed (from left to right) decision trees used in decision theory for determining optimal courses of action [78].

evaluated by processing another part of the data, set aside for testing. If the resulting performance is found acceptable, the ML workflow is complete, otherwise one may need to repeat some of the previous steps.

Suppose that we had followed the above procedure to set up a trained ML model for the problem of forensic identification (as defined in Section 2.1). Would it be of any practical use? Would it be sensible to think that it could be used to process real-case inputs (whose labels are *unknown*) and provide *categorical* conclusions for those inputs (i.e., to label them)? This paper argues that there are (at least) two fundamental problems with the idea that the standard ML setting could serve as a template for forensic identification (conclusions).

The first is a two-fold design problem regarding the question being addressed, which reaches too far. One the one hand, recall that a core design feature of a ML model – defined in the sense outlined above – is to output *direct* identification conclusions. Yet, as clarified in Section 2.1, identification (open set) is a practical impossibility. Arguably, a ML setting with the built-in type of conclusion 'identification' attempts to achieve the impossible, i.e. some sort of – using Salmon's words – "epistemological magic." [85, at p. 66] Thus, using such a system would amount to trying "to perform '*real* magic."' [85, at p. 66] On the other hand, even if one were to accept a less-than-perfect form of identification, it would still – from a legal point of view – violate the procedural scope of action attributed to expert evidence. That scope excludes decisions.

The second problem is methodological and has to do with the data-centrism of the standard ML scheme. As is clear from the description given in Section 4.2, the standard ML scheme is based on the idea of 'cranking out' an answer (output) by relying on (observed) data *only*. This has been called and criticised as the "radical empiricist agenda for machine learning research" [76, at p. 79]. An exclusively data-driven scheme will produce outputs that fall short of the fundamental ingredients that define forensic identification, namely value judgments (i.e., preferences among decision consequences) and prior probabilities referring to other, non-scientific evidence available in the instant case. By way of example, suppose one wishes to use a RF model for forensic identification: such a model would process a given input through a forest of trees, constructed using training data only, and output a conclusion based on majority vote. Such a procedure gives no regard to our relative aversion against a false identification as compared to a missed identification (i.e., preferences among decision consequences), nor can it take into account or be readily combined[44] with any other information that one may have regarding the competing propositions of interest. Yet, both of these aspects are essential.

In view of these stumbling blocks, one might argue that all that needs to be done is 'fix' the standard ML scheme by modifying, for example, the nature of the output (i.e., abstaining from making categorical statements).[45] This, however, would bring us back to what conventional statistical (learning) procedures already do. These produce, as outputs, value of evidence expressions (in terms of likelihood ratios). In addition, methods exist to assess the performance such procedures using data similar to those used in the ML scheme [e.g., 70,79,80].

Yet another objection to this paper's critique of the standard ML scheme might be that preferences among decision consequences could be programmed into the procedure. But this, too, would neither be novel nor solve the problem. First, because preferences in decision analysis are already covered by decision-theoretic accounts of forensic identification [e.g., 12,16].[46] Second, the point of decision-theoretic accounts is not to actually *defer* decision-making authority to an abstract device (machine). Quite to the contrary, in legal applications, decision-making is operated by humans, not by machines – set aside some (low-level) tasks, such as triaging and *preliminary* classifications during investigation. Decision theory is merely one (other) way to articulate what is at stake in forensic identification [28], but the theory does neither tell one what one's beliefs nor one's preferences should be; only how to logically combine these two ingredients.

It is not contested here, however, that there are some tasks for which ML applications may be deployed in forensic science, such as classification in the broader sense of assigning an object or item to a particular category of items (i.e., rather than individualisation in the sense of reducing a set of possible sources to a single member). An example are systems designed to help recognise images with illegal content [e.g., 105] to assist practitioners in dealing with large quantities of data under time constraints. But, as is clear, even for domain-specific classification tasks, users might wish to remain in charge of making the 'final calls', by manually inspecting a system's item labelling (categorisation). The reason for this is that conclusions in classification may be of legal relevance (e.g., the classification of a given item as an illegal drug [20]) and have procedural implications for individuals.

In summary, one can see that the architecture of the standard ML setup cannot capture the essence of forensic identification. ML focuses on 'learning' the associations between inputs and their respective labels, the success of which may be empirically investigated using testing data. Yet, when it comes to treating a new (i.e., real-world) case *beyond* the training and testing data, the labelling of an input is an operation that requires *more* – as was pointed out – than what may have been learned from past data *only*, whatever the quantity of those data. Stated otherwise, forensic identification in the instant case cannot be 'learned' in the way standard ML procedures operate. Hence, publications that suggest that forensic identification could be dealt with as a problem of standard ML represent yet one other form in which unwarranted identification claims persist in forensic science literature.

5. Conclusions

Forensic identification – in the sense of individualisation [64] – is part and parcel of the way forensic scientists operate and widely regarded as an asset to the criminal justice system. And yet, forensic identification has a discomforting dark side, due to its shaky theoretical foundations. Practitioners often circumvent this issue by arguing, somewhat circularly, that identification "works" because they can demonstrate the ability to reliably make identifications *under well--defined*, but often idealistic, conditions. While some forensic scientists have chosen to further develop their reporting practice, away from categorical identification statements toward value of evidence expressions [e.g., 73], many other forensic scientists have not [96]. This is much to the frustration of consumers of expert information who are concerned about the potential of overstatements and contradictions in terms.[47] Changes in policy and practice being reputedly slow, and the willingness of some quarters of forensic science to review and innovate their reporting schemes rather limited [e.g., 31], the burden of scrutinising forensic conclusions in casework will continue to remain, at least for the near future, a constant challenge. This burden comes at a cost that not every defendant would be able to afford.

The conclusion, however, that the persistence of identification

[44] See Ref. [97] for a similar argument as to why posterior probabilities are inadequate for reporting on the value of evidence.

[45] As mentioned above, given the vast array of ML approaches, some ML techniques may evade the critiques in this paper, at least partially, but this is not detrimental to the specific examples that have been evoked, and the fundamental problems encountered by those examples, in particular the provision of outputs in the form of categorical identification conclusions.

[46] For an account on the role of decision theory in AI, see e.g. Ref. [56].

[47] An example are statements of "Source identification (i.e., came from the same source)" [106, at p. 1] *without* excluding all other potential sources. Concerns about this type of reporting language have been expressed, for example, by policy-making bodies in the U.S. (see e.g. references and discussion presented in Sections 1 and 3.2.1).

claims is merely a problem of forensic *practice*, would be short-sighted. As has been argued in this paper, forensic science literature, too, takes its share in perpetuating problematic forensic identification claims. To better understand this phenomenon and its adverse consequences, this paper has exposed three publication strands through which forensic identifications claims are commonly made. These strands have been called descriptivism, diagnosticism and machinism, with the latter being a shorthand term for the now increasingly fashionable approaches based on machine learning. The problem of these publication strands is that, rather than acknowledging that identification is a conclusion that goes beyond what science can provide [95], the contrary is taken as a premise. What is more, as has been shown with reference to Ref. [90], identification conclusions are prone to raise a host of further problems, such as the unwarranted carrying over of source conclusions to ultimate issues and the confusing use of "match" terminology and probability concepts. Thus, problematic identification claims do not manifest themselves in isolation, but arise as a stack of convoluted problems. And yet, it is exasperating to note that these problems are neither new nor unavoidable.

This paper has pinpointed to selected publications to demonstrate that the continuing use of the classic identification paradigm in current forensic science literature, and the persistence of associated problems therein, is more than a mere theoretical consideration. It rather permeates every part of forensic science. The discussed publications were selected, however, for the sole purpose of illustration. They do not represent the main point of the argument. Instead, the primary purpose was to uncover more general properties of the various ways (i.e. publication strands) whereby identification claims are commonly made. The three publication strands discussed in Sections 2 to 4, while not claiming to provide an infallible or exhaustive account of the problem, are an attempt to bettering our understanding of the way in which problematic identification claims arise, which can serve as a first step toward avoiding them.

Thus, on pain of contributing to the persistence of domain-wide pursuits of unwarranted identification claims, forensic science journals should exert greater care in ensuring that publications properly acknowledge the suitability, scope and limitations of research methods used in studies involving the notion of identification. In view of the analysis presented in this paper, more emphasis should be placed on the distinction between, on the one hand, improvements on "tool problems" that use forensic identification merely as an illustrative example and, on the other hand, studies that make no compromise on the characteristics of forensic identification in the first place, and that choose research methods as a function of these fundamental understandings.

Acknowledgments

This research was supported by the Swiss National Science Foundation through Grant No. BSSGI0_155809.

References

[1] Advisory Committee on Evidence Rules, Minutes of the Meeting of May 3, 2019, Washington, DC, 2019. URL https://www.uscourts.gov/sites/default/files/final_-_minutes_of_the_spring_2019_meeting_ of_the_evidence_rules_committee_0.pdf.
[2] C.G.G. Aitken, F. Taroni, S. Bozza, Statistics and the Evaluation of Evidence for Forensic Scientists, third ed., John Wiley & Sons, Chichester, 2020.
[3] R.J. Allen, Rationality, algorithms and juridical proof: a preliminary inquiry, Int. J. Evid. Proof, Special Issue 1 (1997) 254–275.
[4] R.J. Allen, The nature of juridical proof: probability as a tool in plausible reasoning, Int. J. Evid. Proof, Special Issue 21 (2017) 133–142.
[5] R.J. Allen, M.S. Pardo, The problematic value of mathematical models of evidence, J. Leg. Stud. 36 (2007) 107–140.
[6] R.J. Allen, A. Stein, Evidence, probability, and the burden of proof, Ariz. Law Rev. 55 (2003) 557–602.
[7] D. Banks, Learning, in: K. Frankish, W.M. Ramsey (Eds.), The Cambridge Handbook of Artificial Intelligence, Cambridge University Press, Cambridge, 2014, pp. 151–167.
[8] A. Biedermann, Letter to the Editor: commentary on "Is it possible to predict the origin of epithelial cells? – a comparison of secondary transfer of skin epithelial cells versus vaginal mucous membrane cells by direct contact, M.M. Bouzga et al., Science & Justice, Sci. Justice 60 (2020) 201–203, https://doi.org/10.1016/j.scijus.2020.02.003".
[9] A. Biedermann, K. Kotsoglou, Forensic science and the principle of excluded middle: "inconclusive" decisions and the structure of error rate studies, Forensic Sci. Int.: Synergy 3 (2021), 100147.
[10] A. Biedermann, J. Vuille, Understanding the logic of forensic identification decisions (without numbers). sui-generis, 2018, pp. 397–413.
[11] A. Biedermann, J. Vuille, The decisional nature of probability and plausibility assessments in juridical evidence and proof, Int. Comment. Evid. 16 (2018) 1–30.
[12] A. Biedermann, S. Bozza, F. Taroni, Decision theoretic properties of forensic identification: underlying logic and argumentative implications, Forensic Sci. Int. 177 (2008) 120–132.
[13] A. Biedermann, P. Garbolino, F. Taroni, The subjectivist interpretation of probability and the problem of individualisation in forensic science, Sci. Justice 53 (2013) 192–200.
[14] A. Biedermann, F. Taroni, C. Aitken, Liberties and constraints of the normative approach to evaluation and decision in forensic science: a discussion towards overcoming some common misconceptions, Law Prob. Risk 13 (2014) 181–191.
[15] A. Biedermann, S. Bozza, F. Taroni, Prediction in forensic science: a critical examination of common understandings, Front. Psychol. 6 (1–4) (2015).
[16] A. Biedermann, S. Bozza, F. Taroni, The decisionalization of individualization, Forensic Sci. Int. 266 (2016) 29–38.
[17] A. Biedermann, S. Bozza, F. Taroni, C. Aitken, The consequences of understanding expert probability reporting as a decision, in: Science & Justice, Special Issue on Measuring and Reporting the Precision of Forensic Likelihood Ratios 57, 2017, pp. 80–85.
[18] A. Blandino, G. Travaini, A. Rifiorito, M.A. Piga, M.B. Casali, Prediction model for autopsy diagnosis of driver and front passenger in fatal road traffic collisions, Forensic Sci. Int. 324 (2021), 110853.
[19] M.A. Boden, GOFAI, in: K. Frankish, W.M. Ramsey (Eds.), The Cambridge Handbook of Artificial Intelligence, Cambridge University Press, Cambridge, 2014, pp. 89–107.
[20] S. Bozza, J. Broséus, P. Esseiva, F. Taroni, Bayesian classification criterion for forensic multivariate data, Forensic Sci. Int. 244 (2014) 295–301.
[21] J.S. Buckleton, J.-A. Bright, D. Taylor, Forensic DNA Evidence Interpretation, second ed., CRC Press, Boca Raton, FL, 2016.
[22] J. Butler, H. Iyer, R. Press, M.K. Taylor, P.M. Vallone, S. Willis, DNA Mixture Interpretation: A NIST Scientific Foundation Review (NISTIR 8351-DRAFT), National Institute of Standards and Technology, Gaithersburg, 2021.
[23] A. Carriquiry, H. Hofmann, X.H. Tai, S. VanderPlas, Machine learning in forensic applications, Significance 16 (2019) 29–35.
[24] A. Casale, J. Dettman, Composite machine learning algorithm for material sourcing, J. Forensic Sci. 65 (2020) 1458–1464.
[25] C. Champod, Identification/individualisation, overview and meaning of ID, in: J. H. Siegel, P.J. Saukko, G.C. Knupfer (Eds.), Encyclopedia of Forensic Science, Academic Press, San Diego, 2000, pp. 1077–1084.
[26] C. Champod, I.W. Evett, Interpretation, a personal odyssey, in: NIST 2017 Technical Colloquium on Weight of Evidence, NIST, Gaithersburg, MD, 2017, June, pp. 27–29.
[27] C. Champod, I.W. Evett, G. Jackson, J. Birkett, Comments on the scale of conclusions proposed by the ad hoc committee of the ENFSI Marks Working Group, in: Information Bulletin for Shoeprint/Toolmark Examiners 6, 2000, pp. 11–18, 3.
[28] C. Champod, C. Lennard, P. Margot, M. Stoilovic, Fingerprints and Other Ridge Skin Impressions, second ed., CRC Press, Boca Raton, 2016.
[29] S.-L. Chiam, D. Higgins, K. Colyvas, M. Page, J. Taylor, Interpretation, confidence and application of the standardised terms: identified, probable, possible, exclude and insufficient in forensic odontology identification, Sci. Justice 61 (2021) 426–434.
[30] 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.
[31] S.A. Cole, Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States, Law Prob. Risk 13 (2014) 117–150.
[32] 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.
[33] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A hierarchy of propositions: deciding which level to address in casework, Sci. Justice 38 (1998) 231–239.
[34] C. Dahlman, Determining the base rate of guilt, Law Probab. Risk 17 (2018) 15–28.
[35] B. de Finetti, Does it make sense to speak of 'good probability appraisers'? in: I. J. Good (Ed.), The Scientist Speculates: an Anthology of Partly-Baked Ideas Basic Books, New York, 1962, pp. 357–364.
[36] B. de Finetti, Philosophical Lectures on Probability, Collected, Edited, and Annotated by Alberto Mura, Synth. Libr. 340 (2008). Springer, New York.
[37] B. de Finetti, Theory of Probability, A Critical Introductory Treatment, reprint edition, John Wiley & Sons, Chichester, 2017.
[38] I. Dennis, The Law of Evidence, sixth ed., Sweet & Maxwell, London, 2017.

[39] D. Dessimoz, C. Champod, Linkages between biometrics and forensic science, in: A.K. Jain, P. Flynn, A.A. Ross (Eds.), Handbook of Biometrics, 2008, pp. 425–459. Springer, New York.
[40] G. Edmond, M.B. Thompson, J.M. Tangen, A guide to interpreting forensic testimony: scientific approaches to fingerprint evidence, Law Probab. Risk 13 (2014) 1–25.
[41] European Network of Forensic Science Institutes (ENFSI), Vision of the European Forensic Science Area 2030: Improving the Reliability and Validity of Forensic Science and Fostering the Implementation of Emerging Technologies, 2021.
[42] I.W. Evett, A quantitative theory for interpreting transfer evidence in criminal cases, Appl. Stat. 33 (1984) 25–32.
[43] I.W. Evett, Interpretation: a personal odyssey, in: C.G.G. Aitken, D.A. Stoney (Eds.), The Use of Statistics in Forensic Science, 1991, pp. 9–22. Ellis Horwood, New York.
[44] I.W. Evett, Establishing the evidential value of a small quantity of material found at a crime scene, J. Forensic Sci. Soc. 33 (1993) 83–86.
[45] I.W. Evett, Avoiding the transposed conditional, Sci. Justice 35 (1995) 127–131.
[46] I.W. Evett, J.A. Lambert, J.S. Buckleton, A Bayesian approach to interpreting footwear marks in forensic casework, Sci. Justice 38 (1998) 241–247.
[47] I.W. Evett, C.E.H. Berger, J.S. Buckleton, C. Champod, G. Jackson, Finding the way forward for forensic science in the US – a commentary on the PCAST report, Forensic Sci. Int. 278 (2017) 16–23.
[48] G. Fournier, F. Savall, A. Galibourg, L. Gély, N. Telmon, D. Maret, Three-dimensional analysis of bitemarks: a validation study using an intraoral scanner, Forensic Sci. Int. 309 (2020), 110198.
[49] R.D. Friedman, Answering the Bayesioskeptical challenge, Int. J. Evid. Proof, Special Issue 1 (1997) 276–291.
[50] P. Garbolino, F. Taroni, Evaluation of scientific evidence using Bayesian networks, Forensic Sci. Int. 125 (2002) 149–155.
[51] P. Gill, T. Hicks, J.M. Butler, E. Connolly, L. Gusmão, B. Kokshoorn, N. Morling, R.H.A. van Oorschot, W. Parson, M. Prinz, P.M. Schneider, T. Sijen, D. Taylor, DNA Commission of the International Society for Forensic Genetics: assessing the value of forensic biological evidence – guidelines highlighting the importance of propositions. Part I: evaluation of DNA profiling comparisons given (sub-) source propositions, Forensic Sci. Int.: Genetics 36 (2018) 189–202.
[52] S. Gittelson, C.E.H. Berger, G. Jackson, I.W. Evett, C. Champod, B. Robertson, J. M. Curran, D. Taylor, B.S. Weir, M.D. Coble, J.S. Buckleton, A response to "Likelihood ratio as weight of evidence: a closer look" by Lund and Iyer, Forensic Sci. Int. 288 (2018) e15–e19.
[53] L. Han, W. Li, Y. Hu, H. Zhang, J. Ma, K. Ma, B. Xiao, G. Fei, Y. Zeng, L. Tian, L. Chen, Model for the prediction of mechanical asphyxia as the cause of death based on four biological indexes in human cardiac tissue, Sci. Justice 61 (2021) 221–226.
[54] E. Hare, H. Hofmann, A. Carriquiry, Algorithmic approaches to match degraded land impressions, Law Probab. Risk 16 (2017) 203–221.
[55] B. Hartung, D. Rauschning, H. Schwender, S. Ritz-Timme, A simple approach to use hand vein patterns as a tool for identification, Forensic Sci. Int. 307 (2020), 110115.
[56] E.J. Horvitz, J.S. Breese, M. Henrion, Decision theory in expert systems and artificial intelligence, Int. J. Approx. Reason. (1988) 247–302. Special Issue on Uncertainty in Artificial Intelligence.
[57] R.A. Howard, A.E. Abbas, Foundations of Decision Analysis, Pearson, Essex, 2016.
[58] INTERPOL., INTERPOL Disaster Victim Identification Guide, 2018. Lyon.
[59] INTERPOL., INTERPOL Disaster Victim Identification Guide, Annexure 12, Methods of identification, Lyon, 2018.
[60] J. Kaplan, Decision theory and the factfinding process, Stanford Law Rev. 20 (1968) 1065–1092.
[61] M.-A. Katsara, W. Branicki, S. Walsh, M. Kayser, M. Nothnagel, Evaluation of supervised machine-learning methods for predicting appearance traits from DNA, Forensic Sci. Int.: Genetics 53 (2021), 102507.
[62] D.H. Kaye, Apples and oranges: confidence coefficients and the burden of persuasion, Cornell Law Rev. 73 (1987) 54–77.
[63] D.H. Kaye, Clarifying the burden of persuasion: what Bayesian decision rules do and do not do, Int. J. Evid. Proof 3 (1999) 1–29.
[64] P.L. Kirk, The ontogeny of criminalistics, J. Crim. law, Criminol. Police Sci. 54 (1963) 235–238.
[65] J.J. Koehler, S. Liu, Fingerprint error rate on close non-matches, J. Forensic Sci. 66 (2021) 129–134.
[66] F. Lad, Operational Subjective Statistical Methods : a Mathematical, Philosophical, and Historical Introduction, John Wiley & Sons, New York, 1996.
[67] D.V. Lindley, Probability, in: C.G.G. Aitken, D.A. Stoney (Eds.), The Use of Statistics in Forensic Science, 1991, pp. 27–50. Ellis Horwood, New York.
[68] D.V. Lindley, Understanding Uncertainty, John Wiley & Sons, Hoboken, 2006.
[69] D.V. Lindley, Foreword, in: B. de Finetti (Ed.), Theory of Probability, A Critical Introductory Treatment, Reprint edition, John Wiley & Sons, Chichester, 2017.
[70] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. 276 (2017) 142–153.
[71] M. Mitchell, Artificial Intelligence, A Guide for Thinking Humans, Pelican Books, London, 2019.
[72] G.S. Morrison, D.H. Kaye, D.J. Balding, D. Taylor, P. Dawid, C.G.G. Aitken, S. Gittelson, G. Zadora, B. Robertson, S. Willis, S. Pope, M. Neil, K.A. Martire, A. Hepler, R.D. Gill, A. Jamieson, J. de Zoete, R.B. Ostrum, A. Caliebe, A comment on the PCAST report: skip the "match"/"non-match" stage, Forensic Sci. Int. 272 (2017) e7–e9.
[73] G.S. Morrison, E. Enzinger, V. Huges, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W.C. Thompson, D. van der Vloed, R.J.F. Ypma, C. Zhang, A. Anonymous, B. Anonymous, Consensus on validation of forensic voice comparison, Sci. Justice 61 (2021) 299–309.
[74] K.P. Murphy, Machine Learning: a Probabilistic Perspective, MIT Press, Cambridge, 2012.
[75] 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.
[76] J. Pearl, Radical empiricism and machine learning research, J. Causal Inference 9 (2021) 78–82.
[77] President's Council of Advisors on Science and Technology (PCAST), Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, 2016. Washington, D.C.
[78] H. Raiffa, Decision Analysis, Introductory Lectures on Choices under Uncertainty, Addison-Wesley, Reading, Massachusetts, 1968.
[79] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int. 230 (2013) 156–169.
[80] D. Ramos, D. Meuwly, R. Haraksim, C.E.H. Berger, Validation of forensic automatic likelihood ratio methods, in: D.L. Banks, K. Kafadar, D.H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC Press, Boca Raton, 2021, pp. 143–163.
[81] B. Robertson, G.A. Vignaux, Interpreting Evidence. Evaluating Forensic Science in the Courtroom, John Wiley & Sons, Chichester, 1995.
[82] B. Robertson, G.A. Vignaux, C.E.H. Berger, Interpreting Evidence. Evaluating Forensic Science in the Courtroom, second ed., John Wiley & Sons, Chichester, 2016.
[83] M.J. Saks, J.J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (2005) 892–895.
[84] M.J. Saks, et al., Forensic bitemark identification: weak foundations, exaggerated claims, J. Law Biosci. 3 (2016) 538–575.
[85] W.C. Salmon, The Foundations of Scientific Inference, University of Pittsburgh Press, Pittsburgh, PA, 1966.
[86] S. Shalev-Shwartz, S. Ben-David (Eds.), Understanding Machine Learning, from Theory to Algorithms, Cambridge University Press, Cambridge, 2014.
[87] B. Shinkins, M. Thompson, S. Mallett, R. Perera, Diagnostic accuracy studies: how to report and analyse inconclusive test results, BMJ 346 (2013) f2778.
[88] 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.
[89] D.L. Simel, D.B. Matchar, J.N. Feussner, Diagnostic tests are not always black or white: or, all that glitters are is not [a] gold [standard], J. Clin. Epidemiol. 44 (1991) 967–971.
[90] A.M. Smith, T.M.S. Neal, The distinction between discriminability and reliability in forensic science, Sci. Justice 61 (2021) 319–331.
[91] D.A. Stoney, Evaluation of associative evidence: choosing the relevant question, J. Forensic Sci. Soc. 24 (1984) 473–482.
[92] D.A. Stoney, Transfer evidence, in: C.G.G. Aitken, D.A. Stoney (Eds.), The Use of Statistics in Forensic Science, Ellis Horwood, New York, 1991, pp. 107–138.
[93] D.A. Stoney, What made us ever think we could individualize using statistics? J. Forensic Sci. Soc. 31 (1991) 197–199.
[94] D.A. Stoney, Relaxation of the assumption of relevance and an application to one-trace and two-trace problems, J. Forensic Sci. Soc. 34 (1994) 17–21.
[95] D.A. Stoney, Discussion on the paper by Neumann, Evett and Skerrett, J. Roy. Stat. Soc. 175 (2012) 399–400.
[96] H.J. Swofford, S.A. Cole, V. King, Mt. Everest – we are going to lose many: a survey of fingerprint examiners' attitudes towards probabilistic reporting, Law Probab. Risk 19 (2021) 255–291.
[97] F. Taroni, A. Biedermann, Inadequacies of posterior probabilities for the assessment of scientific evidence, Law Probab. Risk 4 (2005) 89–114.
[98] F. Taroni, A. Biedermann, P. Garbolino, C.G.G. Aitken, A general approach to Bayesian networks for the interpretation of evidence, Forensic Sci. Int. 139 (2004) 5–16.
[99] F. Taroni, P. Garbolino, A. Biedermann, C. Aitken, S. Bozza, Reconciliation of subjective probabilities and frequencies in forensic science, Law Probab. Risk 17 (2018) 243–262.
[100] F. Taroni, S. Bozza, A. Biedermann, Decision theory, in: D.L. Banks, K. Kafadar, D. H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC Press, Boca Raton, 2021, pp. 103–130.
[101] D. Taylor, J. Buckleton, I. Evett, Testing likelihood ratios produced from complex DNA profiles, Forensic Sci. Int.: Genetics 16 (2015) 165–171.
[102] W.C. Thompson, How should forensic scientists present source conclusions? Seton Hall Rev. 48 (2018) 773–813.
[103] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy, Law Hum. Behav. 11 (1987) 167–187.
[104] W.C. Thompson, F. Taroni, C.G.G. Aitken, How the probability of a false positive affects the value of DNA evidence, J. Forensic Sci. 48 (2003) 47–54.
[105] A. Ulges, A. Stahl, Automatic detection of child pornography using color visual words, in: In 2011 IEEE International Conference on Multimedia and Expo, 2011, pp. 1–6.
[106] U.S. Department of Justice, Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline, Vers. 8.15.20, Available online at, https://www.justice.gov/olp/page/file/1284786/download, 2020.
[107] M.S. Veldhuis, S. Ariëns, R.J.F. Ypma, T. Abeel, C.C.G. Benschop, Explainable artificial intelligence in forensics: realistic explanations for number of contributor predictions of DNA profiles, Forensic Sci. Int.: Genetics 56 (2022), 102632.

[108] R.B. Weimer, D.M. Wright, T. Hodgins, Paints and polymers, in: V.J. Desiderio, C. E. Taylor, N.N. Daéid (Eds.), Handbook of Trace Evidence Analysis, Wiley, Hoboken, NJ, 2021, pp. 157–218.
[109] S.M. Willis, L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, A. Nordgaard, C.E.H. Berger, M.J. Sjerps, J.J. Lucena-Molina, G. Zadora, 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.
[110] A.C. Zamora, S.D. Tallman, The role of diffuse idiopathic skeletal hyperostosis (DISH) in positive identification, J. Forensic Sci. (2021) (in press).