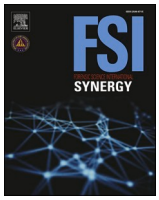
Forensic Science International: Synergy 9 (2024) 100479

Contents lists available at ScienceDirect

Forensic Science International: Synergy

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



Polygraph-based deception detection and Machine Learning. Combining the Worst of Both Worlds?

ARTICLE INFO

Keywords
Polygraph screening
Machine learning
Research methodology
Legal process
Classification
Inference structures

ABSTRACT

At a time when developments in computational approaches, often associated with the now much-vaunted terms Machine Learning (ML) and Artificial Intelligence (AI), face increasing challenges in terms of fairness, transparency and accountability, the temptation for researchers to apply mainstream ML methods to virtually any type of data seems to remain irresistible. In this paper we critically examine a recent proposal to apply ML to polygraph screening results (where human interviewers have made a conclusion about deception), which raises several questions about the purpose and the design of the research, particularly given the vacuous scientific status of polygraph-based procedures themselves. We argue that in high-stake environments such as criminal justice and employment practice, where fundamental rights and principles of justice are at stake, the legal and ethical considerations for scientific research are heightened. Specifically, we argue that the combination of ambiguously labelled data and ad hoc ML models does not meet this requirement. Worse, such research can inappropriately legitimise otherwise scientifically invalid, indeed pseudo-scientific methods such as polygraph-based deception detection, especially when presented in a reputable scientific journal. We conclude that methodological concerns, such as those highlighted in this paper, should be addressed *before* research can be said to contribute to resolving any of the fundamental validity issues that underlie methods and techniques used in legal proceedings.

1. Introduction

There is no shortage of new ideas in the persistent, often desperate, human effort to detect deception in a way that is reliable and free of adhockeries. The ancient dream of separating true from false statements has shaped mythology, philosophy, literature, and fiction [1]. Mary Poppins, for example, immediately upon her return to the Banks family, inquired about the children's behaviour. Her shortcut to the children's inner world is an oral "thermometer" informing her reliably that, for example, Michael has been "careless, thoughtless and untidy" ([2], p. 156; [3]). This amuses adults and children alike because they can distinguish fiction from science, i.e. the difference between literature and reality. But this clear distinction is beginning to erode. It has become commonplace to ask what could be achieved if we used state-of-the-art technology, especially machine learning, to compensate for human shortcomings such as inferential fallibility. Modern technology could offer solutions to the age-old problem of discovering truth, a common idea, in virtually every area of human activity, including forensic science.

The potential for algorithmic approaches in forensic science is considerable [4]. At the comparison stage, for example, when assessing the degree of (dis)similarity between two compared items, marks or traces at the level of their features, variations in the results of the same and different human examiners could be attenuated or even avoided by the use of machine-based or -assisted measurement procedures. Mattijssen et al. [5] present an example of this in the area of feature-comparison as applied to the examination of firearms.

Once observations from a particular comparison are available, algorithmic approaches can also assist in assessing the probative value of those observations. In this context, assessing the probative value means quantifying the extent to which the observations and measurements made during a comparison support a proposition (e.g., "the striation mark comes from the seized tool") over a relevant alternative (e.g., "the striation mark comes from an unknown tool") (e.g. Ref. [6]). Because this task may involve complex computations, there is potential for computational methods to perform such operations with unprecedented speed and reliability that vastly exceed the capacity of the unaided human mind. An illustrative example of this is the emergence of probabilistic genotyping systems for evaluating complex DNA profiling results, especially DNA mixtures [7].

While these developments are laudable, progress in this area is fragile. As we will show in this paper, current research, practice, and publication activities are prone to misconceptions and susceptible to careless research methodologies that can undermine much-needed trust in the use of science in legal proceedings. Threats come from a variety of sources and can cumulate, complicating the problem. Consider, for example, the need for terminological and conceptual clarity regarding purpose and methods. In many applications, these aspects are far from clear. For example, it is no secret that many data-related (research) activities that were broadly and accurately described as statistics a

https://doi.org/10.1016/j.fsisyn.2024.100479
Received 10 April 2024; Received in revised form 10 May 2024; Accepted 13 May 2024
Available online 13 June 2024

decade ago are now increasingly associated with the fashionable terms *data science*, *artificial intelligence* (AI), and *machine learning* (ML). However, many AI and ML approaches are either essentially statistical in nature or, on the contrary, are brute force methods that lack the rigour and explainability of sound statistical methods. Moreover, it is not always the case that rigour drives the research purpose, design, and method selection. Instead, it is often the case that convenience data and ad hoc methods are chosen first, and the purpose is constructed *ex-post* on the basis of the data and methods chosen for data analysis. One such example is presented and discussed in this paper.

These introductory considerations are not a mere caricature of the literature on the application of AI and ML to problems in legal proceedings. As we will show, the problems raised above are real and even appear in papers published in prestigious journals. In particular, we aim to show in this paper that ML approaches applied to problems in legal (investigative) proceedings can be flawed in the sense that they openly ignore the goal of working towards ground truth and that they are used to seemingly legitimise other, scientifically discredited methods. This inevitably has a negative impact on the reputation of ML itself and, more importantly, raises ethical concerns.

Realistically, in a single paper we cannot provide a quantitative survey to characterise the extent of this problem, nor can we hope to remedy it. To suggest otherwise would be tantamount to attempting to overcome Brandolini's principle [8].[1] However, despite the fact that we can only reveal the tip of an iceberg, the critical exposure of problematic research is more important than ever, especially at a time that some have called "a dangerous moment for science and the public" [9,10].

As a running illustrative example, we will take a closer look at an article published in the *Nature* journal *Scientific Reports*. The article in question attempts to develop a "second-opinion tool" for classical polygraph screening aimed at detecting deception [11]. Polygraph screening may seem far removed from the interests of the readership of *FSI Synergy*, but there are several reasons why it is a good topic for discussion. *First*, the technology is widely used in law enforcement and many professional settings. This means that despite longstanding and fundamental criticism, members of the public are potentially exposed to this technology. *Second*, there is probably no other method in legal proceedings that so prominently displays the notion of truthfulness and the aspiration to truth-conduciveness, while at the same time relying on such controversial scientific foundations, as polygraph-based procedures. *Third*, deception detection can be seen as a template for any forensic science method that seeks to help discriminate between competing propositions regarding, for example, the source of evidential material, or alleged activities.

These considerations raise the interesting question of what, if anything, ML can contribute to the scrutiny of polygraph-based procedures and to the meaningful application of computational methods to problems of legal interest more generally. Polygraph screening also provides an opportunity to discuss the proper understanding of the concepts of ground *truth* and *error*, both of which are fraught with misunderstanding in forensic science [12,13].

This paper is structured as follows. *Section* 2 provides a brief introduction to the scientific status and current use of the polygraph in various areas of practice. *Section* 3 presents general elements of performance evaluation for forensic (science) methods. These are contrasted in *Section* 4 with elements of standard ML methodology and, in particular, the methodological choices made by Asonov et al. [11]. In particular, we will critically examine their use of data on deception detection by human examiners based on polygraph screening data. Discussion and conclusions are presented in *Section* 5.

2. Scientific status and current applications of polygraph-based interview procedures

The polygraph does not really need much of an introduction. However, it is helpful to provide some background and context. To begin with, it should be noted that the polygraph is not a self-contained lie detection technique, contrary to what is often suggested. The polygraph is only a recording instrument for various physiological reactions of a person (interviewee). In this sense, raw polygraph screening data is uninterpreted. It takes a human interviewer to reach a conclusion *based* on polygraph screening data. Put differently, deception is *inferred* by the interviewer based on polygraph screening data. Thus, there is a difference between physiological measurements on the one hand and conclusions drawn by an interviewer on the other. The latter requires what is commonly referred to as a "paradigm" of questioning and interrogation, of which there are several, such as the Comparison Question Test or the Concealed Information Test (see, e.g. Ref. [14], for a review).

In this paper we do not elaborate on such different application paradigms. We consider particular interrogation paradigms and physiological target measures to be combined methodological packages, and we will occasionally, and for simplicity, refer to such methodological packages as "polygraph" throughout this paper. Rather, we are interested in discussing how pairs of polygraph screening data and human-assigned conclusions (so-called "labels"), such as "deception indicated" and "no deception indicated", are used in conjunction with machine learning methods, regardless of the polygraph application paradigm chosen.

More generally, it should also be noted that the ability of polygraph-based interview methods to reveal truthfulness is, to put it mildly, highly controversial. Since the polygraph was first implemented to infer deception, it has been consistently and almost universally invalidated in terms of its applicability in fact-finding processes. This critical position has emerged in academic discourse [15,16] and has been echoed by scientific institutions [17,18] and criminal courts (e.g. *Frye v. United States*, 293 F. 1013, D.C. Cir. 1923) that have addressed the issue. Similarly, the approach has fared no better when investigated by the military [19].

The main criticism at a functional level is that an interviewee's physiological response may be due to a variety of factors, not necessarily deception ([18], p. 78; [16], p. 29). Historically, this has not always been properly appreciated. On the contrary, the development of the polygraph and its use to infer deception was influenced by a now discredited and outdated paradigm in psychology based on the false assumption that physiological processes reflect mental processes in an identifiable way, known as psychophysical parallelism [3].

However, these fundamental limitations have not prevented polygraph-based deception detection from being deployed in legal systems. In fact, the polygraph is considered useful as an interrogation aid by its very ability to give the subject the *impression* that it is actually working. In other words, once the subject believes in the infallibility of the polygraph-based procedure, the *bogus pipeline* effect takes place. That is, simply because the interview is conducted under the false assumption that the polygraph (*a*) operates independently[2] and (*b*) will ultimately reflect the person's true attitude, and because the interviewee does not want to be second-guessed by the machine, he or she will feel pressured to disclose unfavourable statements [3,20]. This is similar to horoscopes, which continue to exist largely because (some) people believe in them. There is, however, a striking difference. With horoscopes, everyone is free to skip them or use them in whatever way they choose to make personal decisions. Polygraph-based deception detection, on the other hand, is in many cases compulsory and has real

[1] Broadly speaking, Brandolini's principle states that the amount of energy required to refute problematic output is much greater than that required to produce it.

[2] As mentioned at the beginning of this section, the polygraph is not an independent lie detection technique, but requires a human examiner to reach a conclusion.

consequences for the individual that are decided by third parties.

Given these examples of complications and distortions, it should be clear that the use of the polygraph to detect deception has no place in any process that seeks to meet the demands of modern and rational systems of governance. Unfortunately, reality tells a different story. Suffice it to say that one of the most striking, understudied and underreported developments in the criminal justice system in England and Wales is the increasing use of polygraph-based interrogation techniques. Despite severe criticism from scientific institutions and academic discourse, the legal system in England and Wales is using long discredited polygraph-based procedures to elicit unfavourable and incriminating statements. Statutory law covers the use of polygraph-based procedures as part of the parole process for released offenders, those convicted of terror-related offences, sexual offences and domestic violence. The private sector also makes extensive use of polygraph-based procedures, typically to screen candidates for work environments that require a higher level of security and integrity, such as the financial sector or critical infrastructure (e.g. Ref. [11]).

3. Elements of method performance evaluation in forensic science

The fragile, to say the least, scientific status of the concept of polygraph interviews, understood in the following as a given combination of the use of the polygraph and an interrogation paradigm (Section 2), raises the question of the methodological requirements of scientific research to which polygraph-based research should adhere, and whether ML could ultimately contribute to improving the truth-conduciveness (e.g. sensitivity and specificity) of polygraph-based procedures. Leaving aside purely descriptive or observational research, and concentrating on merely reflecting the state of the art in polygraph interviews, one might be tempted to consider the following two-step procedure: Roughly speaking, start by determining what kind of existing data might be available, and then look at what standard data processing methods are available. These methods could then be applied to the data to help answer questions that fall within the scope of the chosen data processing method(s). The rest would follow standard ML methodology, i.e. evaluate the results of the different data processing methods using established performance metrics, such as error rates, and then factually report the results.

In the remainder of this paper, we argue *against* this idea and use a practical example to illustrate our point. In particular, rather than choosing data and data processing methods first, we insist on the importance of defining overarching goals first, and only then making methodological choices that might serve those goals. The primary goal we have in mind here with respect to methods and techniques in the context of legal proceedings can be generally defined as promoting accuracy in fact-finding (e.g. Ref. [21,22]). That is, in the absence of *knowledge* of the ground truth, the idea is to invoke and rely on methods and techniques that are *thought* to guide one towards reducing certain types of error. Broadly speaking, a viable method or technique is one that has a demonstred level performance in helping people to perform better than they would without formal assistance. While this may sound obvious, it is not always properly appreciated in current research, as we will show.

In forensic science, the notion of ground truth as the relevant reference point is widely accepted. The paradigmatic example of this is the concept of the black-box study as advocated in the PCAST report [23]. Black box studies are studies "in which many examiners render decisions about many independent tests (typically, involving "questioned" samples and one or more "known" samples) and the error rates are determined" ([23], pp. 5–6). Proficiency testing of examiners is also often mentioned in this context. The focus may be on evaluating the performance of a particular method, an examiner, or both, especially when a method cannot be clearly separated from a human operator. The term validation study is another commonly used term in this context. It can be used as a summary term for a wide range of different types of ground truth testing studies.

It is difficult to overstate the importance of known ground truth as a key experimental design feature of studies aimed at evaluating the performance of forensic examiners and/or methods. Meuwly et al. [24], for example, provide a detailed validation guideline for value of evidence methods across trace types based on definitions previously described in Haraksim et al. [25]. They clearly state that "[i]n the validation stage we evaluate the […] method performance using the validation dataset (with a *known ground truth*)" ([24], p. 149, *emphasis added)*. More recently, and along the same lines, Morrison et al. [26] presented principles for validation in the specific area of forensic voice comparison: "The performance of the system is […] assessed by comparing the […] output by the system with the *truth* as to whether they resulted from same-speaker or different-speaker comparisons" ([26], p. 301, *emphasis added*).

As in life in general, the devil lies in the details, especially when the scope of inquiry is extended into the jurisprudential domain, which introduces an unbridgeable gap with respect to the ground truth-focused account of forensic evidence given above. As a preliminary, it is important to recall that the "epistemic" objectives of the criminal process are normatively constituted. This point can be illustrated by reference to the values of legal orders, liberal or authoritarian, as the case may be. Regardless of policy matters and values salient in the respective legal order, factual accuracy, is *not* the *primary* concern of the criminal process (e.g. Ref. [22] for more analysis). Even if the procedural framework of the criminal process aims at minimising certain types of undesirable consequences, empirical reality cannot serve as a reference point in this endeavour. In other words, since the truth value of the (propositional content of the) criminal verdict cannot be determined by external means, the term veritistic cannot be used in this context. From a jurisprudential point of view, therefore, any veritistic approach to evidence is seriously flawed. This does not, however, contradict the above account of ground truth testing in forensic science, which is concerned with aggregate measures of performance established in experiments under controlled conditions and propositions *other* than ultimate issues. In criminal cases, the focus is on the verdict in the particular case, which is not analogous to an exercise in testing under controlled conditions.

4. The problem of data over mind: how polygraph interview data can make a bad case for machine learning

This section examines how the general principles for evaluating the performance of forensic methods outlined in the previous section can be compromised when problems are approached from an overly uncritical standard ML perspective. To this end, it is helpful to introduce some elements of mainstream ML methodology. Section 4.1 provides a brief description of common ML templates. Readers familiar with ML methodologies may wish to skip this section. Because of the ease with which criticisms of ML can be misunderstood [27], it is worth emphasising that Section 4.1 is not an exhaustive account and is therefore not representative of the field of ML *as a whole*. Moreover, our criticism in later sections relates to the specific application we have chosen as a running example. More generally, the point we seek to make is that it is important to expose those aspects and variants of ML that are prone to misapplication in forensic contexts because they reflect poorly on the rest of ML, especially on the meaningful and responsible uses of ML [4]. Section 4.2 discusses the ML methodology used in Asonov et al. [11] as an example of how research can conflict with the principles of performance evaluation of forensic science methods.

4.1. Aspects of data-centric mainstream ML methodology

In informal discussions, ML and artificial intelligence (AI) are often mentioned together, although they are not the same. ML is only one part of the broad AI landscape. We can delineate this landscape along two

dimensions (see e.g. [28] for an overview). One dimension relates to the types of problems or tasks that computational procedures aim to solve, such as perception, reasoning, knowledge, planning, and communication. Another dimension relates to the computational procedure that is used to perform the specific task(s) of interest. Historically, different approaches can be distinguished along this dimension. In the era before the current era of ML, between the 1950s and 1990s, also known as classical AI, a predominant idea was that the tasks to be performed by a machine were fully defined in the code written by human programmers in symbolic form. This approach is well suited for tasks involving probabilistic reasoning and logic, but can be less effective for tasks where humans cannot provide a complete description of how to proceed.

Since around the mid-1990s, the gradually evolving perspective now known as ML has sought to overcome the gaps in human ability to provide a complete specification of the task to be performed by algorithmic systems. ML attempts to do this by presenting the machine with examples of data, known as *training data*, for which the category membership is known. In other words, these are instances where there are measurements of characteristics (features) of an item or event for which we know which category it belongs to. Through several such examples, the so-called learning algorithm tries to find the parameters of a function (in the current example, a classifier) that will later be used to process new items, i.e. measurements of their features (serving as inputs), but for which the category membership (i.e. ground truth label or output) is *not* known. The performance of the learned function, the classifier, is then evaluated using so-called *test data*. The results are summarised using various standard metrics, such as error rates. If the performance is unsatisfactory, the researcher may go back and critically examine the training data (e.g. its quality) and/or the learning algorithm itself. Sometimes the researcher may simply use many different methods on the same training data to see which method leads to the best classification performance.

Note that for the purposes of this discussion we will focus on classification as a common example of supervised learning. Here the term "supervised" refers to the fact that the input and output values for the training data are known. However, this is not the only type of learning. In *un*supervised learning, for example, there are no a priori known class/category assignments. A typical example is data clustering. Finally, we should also mention that classification is only one type of task to which ML is applied. Another common task, which is beyond the scope of our discussion, is regression. It aims to provide a real-valued *output* based on some *input*.

The above description of standard ML methodology is broad and general, but sufficient to illustrate the main ideas. It is important to note that the *learning* step in the methodology, at least in the case of classification, involves training data with *known* ground truth. Moreover, in the performance evaluation step, the output labels of the procedure are compared with the actual ground truth, since factual accuracy is the relevant reference point.

However, complications and controversies can arise when the focus is on details. For example, a recurring controversial issue is the (degree of) transparency and understandability of the procedure at different levels of detail. The question is whether, for a given procedure that processes inputs to outputs (e.g. a classifier), one can reconstruct how the result was produced. In some cases, the procedure amounts to a highly transparent and rigorous *statistical* model (e.g. in the case of *statistical* learning techniques), so that it is actually unnecessary to use the fashionable term "ML" to refer to it. For such procedures, the functioning and results are tractable in the sense that if something does not work as expected, it is possible to inspect the procedure (and its implementation) to search for the reason(s) for the observed behaviour. However, many contemporary ML methods do not have this level of transparency because their functioning is opaque (e.g. Ref. [29]). That is, regardless of whether the procedure works (well) or not, as measured during testing, the researchers would not be able to explain why. Worse, if the procedure works differently than intended, or not well enough, the researchers have no way of discovering the source of the problem(s), except perhaps by trial and error. Needless to say, this is probably one of the least desirable properties of a method intended for use in legal evidence & proof proceedings, which require results based on a sufficiently reliable scientific foundation, not just "turning knobs".

Closely related to the problem of opacity may be the lack of a structural understanding of the problem domain, i.e. a *model* of reality, at least at a qualitative level, which requires elements – i.e. knowledge – beyond the naked data itself. This problem arises when taking the ML methodology outlined above to the point of assuming that all wisdom lies in the data alone. In other words, researchers proceed on the assumption that the computational procedure can develop the ability to perform certain tasks on its own, simply by processing training data. However, the lack of structural knowledge about the problem domain means a lack of insight into the problem of interest. Consequently, when moving from one context to another, for example, the entire training process of the ML procedure may have to be started from scratch.

The above complications should not be surprising from a data-over-mind[3] perspective, which at best seeks to find *associations* between inputs and outputs, without looking further behind the "curtain" to uncover more substantive aspects of the problem of interest. With these elements in mind, we are now ready to critically examine a recent example of the application of data-centric ML methodology to the specific problem of polygraph deception detection, as presented by Asonov et al. [11]. In the next section, we present the ML methodology used by Asonov et al. [11]. We show how it runs counter to the goal of truth-conduciveness and why it raises research ethics concerns.

4.2. Application of ML to polygraph screening/interview data

The running example we consider here is the work reported by Asonov et al. [11], which focuses on "building a second-opinion tool for classical polygraph [screening]".[4] Their study addresses the question of how to improve the review of conclusions drawn by polygraph examiners, who are inevitably prone to error. Specifically, the authors seek to use ML as a novel way to design a computerised device for reviewing polygraph-based conclusions drawn by human examiners.

We will first look at the notion of error as it relates to polygraph screening conclusions (Section 4.2.1), followed by a critical review of the application of the conventional ML template to polygraph interview data (Section 4.2.2). Sections 4.2.3 and 4.2.4 discuss the problem of ambiguously labelled data and the consequences of model-blind machine learning approaches from a structural point of view.

4.2.1. Conclusions of human examiners in polygraph interviews and the notion error

In practice, two types of interview results – i.e. conclusions reached by polygraph interviewers – require review: "deception indicated" (DI) and "no deception indicated" (NDI). Each of these two results can be either accurate or inaccurate, depending on whether or not the interviewee is truthful about a particular subject of interrogation. Before proceeding, however, a general comment on the central notion of error is in order. Asonov et al. [11] argue that "screening errors are not only due to the method, but also due to human (polygraph examiner) errors". We question whether this distinction is helpful in practice, as the polygraph interviewer is an integral part of the screening process and it is difficult to separate the two.

The whole polygraph procedure, including the conclusion reached by the polygraph interviewer, is in fact *an interrogation tool*, another

[3] We paraphrase here Pearl's expression of "mind over data" [30].

[4] Note that Asonov et al. [11] distinguish between classical polygraph screening, which focuses on measuring aspects such as cardiovascular activity and respiration, and other types of polygraph screening, which are based on other variables measured using video and audio recordings.

Trojan horse that hides the extraction imperative under the veil of technological progress [3]. Its purpose and only potential is not to detect truth but to enable interrogators to extract confessional statements at the expense of rationality and legitimacy (see Ref. [1] for further discussion). Furthermore, what Asonov et al. [11] presumably refer to as the "method" is the procedural component that involves physical measurements of the interrogated person. These measurements may be inaccurate to some degree. However, they are not directly an "error" in the conventional sense, i.e. with respect to the ground truth underlying the subject about which the examinee is being questioned. The reason for this, as explained in Section 2, is that DI and NDI conclusions are drawn *by the polygraph interviewer* on the basis of physiological measurements, not by the device used to make those measurements.

This distinction is important because, at this point, Asonov et al. are putting the technological cart before the evidential horse. The main idea underlying polygraph interviews is one of the most characteristic tenets of an obsolete paradigm of psychology (Introspection), i.e. the so-called psychophysical parallelism, according to which mental processes run parallel to physiological ones [31]. This idea became the linchpin of the polygraph. However, the field did not survive the 1920s because it had too many internal methodological inconsistencies and relied on too many idealisations.

It is also worth noting that the conclusions DI and NDI are not as categorical as one might be inclined to think, based on what the common notion of "lie detection" suggests. Polygraph interviewers do not directly "detect" lies, strictly speaking, but only *indicate* deception, as the terms DI and NDI imply. Therefore, DI and NDI conclusions express nothing more than a subjective, unverifiable suspicion. In practice, however, this subtlety is often ignored because consumers of polygraph interview results often confuse indications of deception with lie detection.

We should also include some critical reflections on the feasibility of error detection in principle. In any field application of polygraph screening, the ground truth is typically unknown, because if we knew the truth value of a proposition, polygraph screening would not be necessary. The complexity of real-life cases adjudicated in the criminal justice system exceeds anything found in the psychologist's laboratory. Real people involved in the criminal justice system have real stakes, complex motivations, and varying recollections of events. Empirical research in this area suffers from a lack of realism, which is a prerequisite for validity. In addition, practical application also faces what empirical researchers call the "base rate problem". As Gudjonsson explains, "[a]t the most basic level we do not know the proportion of suspects interrogated at police stations who are genuinely guilty of the offence of which they are accused." ([32], p. 173; see also [3] for further discussion).

The lack of knowledge of the ground truth means that errors, i.e. a discrepancy between the examiner's conclusion and the ground truth, cannot be detected. At best, a second examiner –human or machine – can review the first examiner's recordings and either agree or disagree with the first examiner's conclusion. Asonov et al. [11] call this a component of quality assurance: "have another examiner review the screening and confirm or disprove [sic] the conclusion of the original examiner" (p. 2). Note, however, that this is somewhat misleadingly phrased since the use of the term "disprove" suggests a strong claim of "falsification," which, as noted above, is impossible in the absence of knowledge of the ground truth. Incidentally, Asonov et al. [11] acknowledge this fallibility by calling a review by a second examiner "not a bulletproof solution" because "the second examiner may make just the same mistake the original examiner did" (p. 2).

The clarifications introduced above set the boundaries for assessing what exactly machine learning with polygraph interview data can and cannot legitimately claim to accomplish. We address this issue in the next section.

4.2.2. Scrutinising the application of the conventional ML template to polygraph interview data

As a preliminary, we should emphasise that the nature and form(at) of the data have a crucial impact on the meaning and appropriateness of the result of a standard ML application (as described in Section 4.1). By the nature of the data, we mean whether the data are the result of carefully designed experiments under controlled conditions or whether they are convenience (field) data.

In the *former* case, i.e. structured experiments, conditions are known and controlled, at least to some extent. In the context of polygraph-based interviews, this condition may be pushed to an unrealistic level. Consider, for example, how researchers simplify case studies to the point of meaninglessness in an attempt to solve the problem of replicability for real events. Subjects in psychological research are typically instructed to imagine committing a mock crime, such as "stealing" something in the room. According to the researchers' intention, this would create an emotional potential with which to experiment. However, this is an experiment with the wrong kind of guinea pigs (see already [33]), for the forensic context is very different from the laboratory conditions outlined above.

In the *latter* case, i.e. convenience (field) data, conditions are more likely to vary from case to case, possibly quite substantially, raising the issue of uncontrolled confounding factors.

By data form(at) we mean, broadly speaking, what exactly the measurements refer to and how the (category) labels are defined. As a simple example, consider a case where the measurements refer to the height of a person, and the (category) labels refer to whether the measured individual is a man or a woman, or more generally, whether the individual is a member of the population (group, class, etc.) 1, 2, 3, etc., depending on the type of classification problem at hand.[5] Thus, as we can see, the extent to which a mainstream ML technique can actually "learn" anything substantial depends crucially on the quality with which the data has been labelled. However, as we will argue later, data alone is not sufficient to ensure a meaningful ML application.

With these observations in mind, let us now turn to the study by Asonov et al. [11]. With regard to their methodology, two important and interrelated aspects need to be mentioned. The first relates to the purpose of the study. The authors point out that their aim is *not* to develop a device that directly outputs statements about the ground truth, i.e. the truthfulness of an interviewee's answers to questions of interest. Instead, the authors are attempting to develop a device that provides a statement about *what a (second) polygraph interviewer would conclude* about the truthfulness of an interviewee's responses (i.e. DI or NDI). For this reason, the authors refer to their device as a "second-opinion tool". The difference between these two purposes is subtle but crucial, as we will see.

Moreover, the term "second-opinion tool" is potentially confusing because a second-opinion is usually understood as a further statement *about* the actual ground truth, not a mere conjecture about what an average examiner would conclude about the ground truth. For example, in the context of weather forecasting, a meaningful second opinion would be a second statement about the future state of the world, such as rain or no rain the next day, *not* about a forecaster's statement (or indication) about rain or no rain the next day. Therefore, we would not characterise Asonov et al.'s [11] device as a proper second-opinion tool, but rather as a *pseudo*-second-opinion tool.

One might be tempted to say that there is no real difference between the two kinds of conclusions distinguished above, or that any difference is immaterial. However, this temptation should be resisted, as will become clear when we turn now to the second aspect related to the methodology used by Asonov et al. [11]: the nature and structure of the data used for ML. The authors state that their data consist of "historical

[5] Recall that, as noted in Section 4.1, classification is only one of several tasks for which ML is used.

data from 2094 field polygraph screening recordings (PSRs) including Deception Indicated (DI) attributes set by the examiners who conducted the screenings" (p. 2). Clearly, this is not data obtained under *controlled conditions* and with known ground truth (i.e. whether the interviewee gave a truthful answer or not), but convenience data labelled with the DI or NDI label of the particular examiner, i.e. something more arbitrary than rational. In practice, using such data for ML purposes allows a model to "learn" the association between the input measurements and the examiner's conclusions (outputs), *not* with respect to the actual ground truth in each case. Put another way, performing ML on such data means mimicking human examiners at their (imperfect) level of performance, rather than designing a device that provides its own conclusion about the relevant ultimate ground truth state.

It is worth pausing for a moment to consider what it means for ML systems to mimic the conclusions of human polygraph examiners. To do this, let us recall the goals of ML. According to at least one viewpoint, the purpose of ML is to develop machines that can perform certain tasks that would otherwise require human skills (intelligence), and ideally perform those tasks better than humans, especially where the tasks are complex and human performance is variable and poor (i.e. prone to error). This, in turn, raises the question of what the task actually is.

In the context of polygraph screening (PS), the obvious primary task would be to infer deception. However, this task would require data labelled with the actual ground truth state (interviewee truthfulness), not just DI and NDI labels as in the convenience data used by Asonov et al. [11]. Thus, the task defined by Asonov et al. [11] can never reach the level of interviewee truthfulness. Instead, as noted in our first point above, the authors' ML tool can at best flag a deviation from the aggregate interviewers' conclusion (i.e. the average wisdom of the crowd). This not only reflects a lack of ambition to improve the truth-conduciveness of practical procedures (i.e. with respect to ground truth), but amounts to a conscious decision to abandon the concept of truth altogether. Asonov et al. thus replace methodological rigour with methodological nihilism. Limiting the scope of ML to mimicking the process of reaching conclusions by polygraph interviewers is tantamount to further enforcing and imposing current PS practice in its imperfect state of development and operation. Such an ML design would also amount to merely making the scientifically discredited practice of PS more economically efficient, if indeed quality assessment (QA) by machine could somehow replace or reduce the need for human resources for the same task. In other words, a research design with data that is deprived of actual ground truth is necessarily incapable of discovering anything substantially new in the effort to draw conclusions about the truthfulness of interviewees' answers.

Regarding the problem of ground truth data, one might object to our critique by arguing that the DI and NDI conclusions are *quasi*-ground truth labels, i.e. acceptable proxies for it. Indeed, Asonov et al. [11] write that "the share of examiner errors [in the data] is minor" (p. 2). However, this objection and assumption is unfounded and speculative. The authors provide no independent evidence that the proportion of incorrect labelling in their data can be considered low or high. Nor do they justify what is *low enough* or what level of error would be acceptable in operational practice. Instead, based on what we know from critical research on PS, we argue that error probabilities can be virtually anything, and even vary across settings, due to the multivariate nature of the problem in the first place. Furthermore, if the error rate were indeed so small as to be negligible, it would beg the question of why PS requires review, which would defeat the purpose of the study.

4.2.3. Illustrating the problem of ML with ambiguously labelled data

The main point made in the previous section was that training a machine on convenience data, where only the interviewer's *presumed* ground truth assignment rather than *actual* ground truth is available, is problematic, regardless of whether the use of such convenience data is intentional rather than the result of a mere practical constraint. While the intricate nature of such a research design may remain difficult for the general readership to understand, we can further illustrate it with some common forensic science examples.

A convenient example is the task of comparing a fingermark of unknown origin found at a crime scene with the reference print of a person of interest, and the assignment of a probative value to the similarities and differences observed during such a comparison. In forensic science, various computational methods have been developed to produce value of evidence assessments for the results of mark-to-print comparisons. As discussed earlier (Section 3), Meuwly et al. [24] proposed guidelines for the validation of such systems. Broadly speaking, such validation focuses on whether a given method produces outputs (quantified expressions of the value of evidence) that are congruent with ground truth. For example, in the case of a comparison where the crime scene mark and the reference print are from the same person, a system should output an evidence value that expresses support for the proposition that the compared items are from the same source, rather than the proposition that they are from different sources. However, there would be no interest in training a system on data with labels that correspond to what examiners *believe* to be the ground truth, rather than the actual ground truth.[6] The reason for this is that we want a system that strives for ground truth as the relevant reference point, not just the level of human examiner performance, which is known to be imperfect, i.e. suboptimal in terms of truth-conduciveness.

Another example, conceptually similar to fingermark examination, is forensic voice comparison. In this area, considerable progress has been made in the development of automated, human-supervised forensic voice comparison systems and their empirical validation under casework conditions. As with fingermark examination, the ground truth is the primary reference point. The performance of forensic voice comparison systems is evaluated against the ground truth, i.e. whether the examined voice recordings are from the same or two different sources (see especially [26]). No one in this field would advocate training and evaluating systems based on speech recordings labelled with what humans *believe* to be the ground truth, because such labels would not only be expected to differ from the ground truth, but would also potentially vary from examiner to examiner. The result would be ambiguously labelled data.

4.2.4. The consequences of model-blind machine learning from a structural point of view

So far, our discussion has focused mainly on the nature of the data and the impact of this aspect on the meaningfulness of the ML process. However, there is more to say about the intricacies of applying the standard ML framework. In particular, it should be kept in mind that the standard ML framework has little to say about the definition of the variables involved, other than generalities such as whether they are discrete or continuous, binary or multi-class, and so on. The ML template can therefore be applied to a wide range of problems, including textbook examples such as classifying papayas as tasty or untasty based on aspects such as softness and colour (e.g. Ref. [34]).[7] While this ease and flexibility of application can be seen as an advantage, it can potentially lead to ignoring aspects of the real world that cannot be extracted from the data alone, regardless of the amount of training data available. Field PS data provide an interesting example for further consideration of this issue from a structural perspective.

In the context of classification problems, it is helpful to consider –

[6] There are rare exceptions in the forensic literature. For example, Dror and Scurich [40] have proposed elevating the conclusion categories of human examiners to ground truth states when calculating error rates, but this proposal has been refuted on grounds of logic [12] and practical feasibility [35]. For further debate on the same topic, see also Biedermann and Kotsoglou [13,36] and Scurich and John [37].

[7] As noted in Section 4.1, we restrict our discussion here to problems of classification and leave aside other tasks, such as regression.

structurally speaking – that the hypothesis (or class) variable *conditions* the variable for which measurements or observations are available. For example, in the case of height measurements for men and women, the distribution of measurements (e.g. the variable "height") is said to depend on, or be conditioned by, the hypothesis variable (man or woman). On the basis of a particular measurement, we can then *infer* something about whether a given measurement is of a man or a woman, without necessarily making a categorical statement about class membership.[8]

Turning now to field PS data, we can ask whether it makes sense to assume that DI and NDI conclusions are categories in the sense explained above. Put differently, the question is whether interviewers' conclusions (of DI and NDI) really *condition* the raw PS results (measurements). This seems implausible in that an interviewer's conclusion is given only *after* the PS measurements have been taken. Furthermore, while the states "deception" (D) and "no deception" (ND) do indeed form a hypothesis variable that represents ground truth, examiners' DI and NDI *conclusions* are *not* a conditioning variable of the same kind.[9] Rather, it is the other way round: an examiner's conclusion (DI or NDI) is *a consequence* of the measurement, i.e. an *ex-post* statement.

To illustrate the above point, we can use DNA analysis as a rough analogy (e.g. Ref. [38]). Here the ground truth (the hypothesis variable) is whether two biological stains come from the same individual or from two different individuals. This proposition logically conditions the variable representing the event whether the stains have corresponding or non-corresponding DNA profiles. Finally, *depending* on whether the stains have corresponding profiles, the scientist will *report* either a correspondence or a non-correspondence. It is important to note that it is *not* the examiner's report of a correspondence that "makes" the DNA profiles correspond to one another, as this would be tantamount to claiming that an effect (variable) "produces" a cause.[10] Instead, the examiner's conclusion depends on whether or not the DNA profiles actually correspond. Similarly, it's not the polygraph examiner's conclusion (DI or NDI) that "causes" the raw PS results (measurements). Instead, the variable representing the examiner's conclusion depends on the PS measurement result. It is only by making the approximation of interpreting the DI and NDI labels assigned by human examiners as proxies for the ground truth labels (D and ND) that the proper diagnostic structure could be restored here. However, as we have argued in Section 4.2.2, this assumption is not warranted.

One might object to the above analysis by claiming that DNA profile comparison and source inference are not the same as polygraph screening and deception detection. This is true in terms of practical details, but not at the level of the *structure of reasoning*. Both applications deal with the general problem of going from a report from a human source (here: an examiner's report) to an assessment of the underlying ground truth via some intermediate variable(s), i.e. a chain of reasoning (e.g. Ref. [39]). This structural analysis is not just a conceptual nicety, but can actually show us how to make research on polygraph interviews potentially more purposeful.

Specifically, the structural analysis of the problem of interest tells us that in a reasoning process in which raw PS results (measurements) are available, an examiner's (mere) opinion of the ground truth, in the form of a DI or NDI report, is irrelevant for inferring anything about the ground truth (i.e. the D or ND states). The reason for this is that, graphically speaking, the measurement results "screen off" the hypothesis variable (D and ND) from the examiner's conclusion (DI and NDI). This brings us back to our main argument: the natural way to build an examiner opinion review system would be to train on data with the actual ground truth labels D and ND, i.e. at the top of the inference chain, rather than on DI and NDI labels, which are at the bottom of the inference chain.

Finally, one could argue that many standard ML methods do not require one to worry about structural details of the kind discussed above. In other words, model-blind methods focus on "learning" the associations between virtually any input and output variables, regardless of how the variables of interest are structurally related to each other. One could even argue that such methods are exactly what is needed, since they are ideal for situations where the task to be performed is not sufficiently understood for humans to formally specify and implement it in a program. However, when relevant structural knowledge about reality *is* available, it is questionable to proceed and base ML on assumptions that are clearly contrary to what that knowledge would suggest.

5. Discussion and conclusions

The development and use of algorithmic approaches in combination with physiological data (including conclusions of human examiners), which purport to provide a second opinion tool for classical polygraph screening, represents a complication nested within a wider set of problems that go beyond pure research settings. Government, industry and society as a whole form a complex, interwoven structure that provides fertile ground for applied research, including research into polygraph-based interview procedures to detect deception. Suffice it to say, for example, that one of the most notable recent developments in the penal system of several jurisdictions, including England and Wales, has been the increased use of polygraph-based interviewing techniques. Previously restricted (by statute) to certain sex offenders released on licence, polygraph-based interviews can now also be imposed on released domestic violence and terrorism offenders, those subject to Terrorism Prevention and Investigatory Measures (TPIMs) and those subject to Sexual Harm Prevention Orders through a recent statutory power to impose positive requirements. In addition, the UK Government is proposing to introduce polygraph-based interview procedures for offenders convicted of murder where there is a risk of them committing a sexual offence on release.

The aforementioned developments in criminal policy highlight, in the strongest possible way, the need for a public debate on the ethics and regulation of the use of technology, particularly in view of the fact that the standards for the use of devices such as the polygraph are not set by the scientific community. In the UK, for example, the design and delivery of the "Basic Polygraph Examiner training" must conform to the rules and standards set by the American Polygraph Association (APA). Although the APA is independent of the executive, it can hardly be described as an *independent* body. On the contrary, it is a trade association with clear commercial interests. While they are more than entitled to promote their business objectives, it remains problematic, indeed scandalous, that *they* set the standards for the assessment, but also for the training and quality control of an intrusive measure [1].

It is tempting and misleading to think or suggest that a lack of scientific validity can be remedied by mere increased computational capacity in general, or by the now fashionable standard ML methods run on convenience data in particular. The fact remains that polygraph-based interview procedures are pseudoscientific for the simple reason that *case-specific* physiological behaviour is more complex than what can be captured by frameworks that purport to identify categorical relationships between psychological concepts and physiological behaviour, i.e. to identify a single cause for externally measurable behaviour, especially since the data include the subjective conclusions of the human examiner as an additional dimension.

[8] We emphasise that our discussion here does not relate to forensic evidence evaluation where it is generally accepted that forensic scientists should *not* opine directly on propositions, but should only assess the value of the findings (observations), i.e. their capacity to help discriminate between competing propositions of interest (e.g. Ref. [6]).

[9] Note that we have underlined the "I" in DI and NDI to highlight that the abbreviation stands for an examiner's *indication*, which is not to be confused with the actual ground truth states D and ND, respectively.

[10] See e.g. Taroni et al. [41] for a graphical representation of this understanding using a Bayesian network.

Building ML models, as Asonov et al. [11] do, with training data for which human interviewers attributed DI and NDI labels in an unverifiable way means that the resulting models at best emulate a generic human interviewer. The actual ground truth as *the* relevant reference point is thus openly dismissed. Worse still, the authors' use of the terms "error (detection)" and "verification" is unfounded and misleading. The term "error" refers to a mismatch with respect to ground truth, and verification requires knowledge of the ground truth – which is absent in Asonov et al.'s data. To suggest otherwise is ethically questionable, to say the least, given the strong interest that polygraph-based procedures continue to enjoy among interested parties.

More generally, the authors' suggestion that the proposed models' aspiration to mere human-level performance is appropriate and sufficient because the intended use of their models is limited to flagging cases for further review by human examiners is shallow. To argue in this way is to suggest that ML is not intended to help practise move towards a level of operation more firmly rooted in factual rectitude, but merely to make a profoundly human black box process more economically efficient, which is ethically reprehensible. At a time when developments in the broader field of AI are facing increasingly harsh challenges over aspects such as fairness and accountability, Asonov et al.'s ML approach is thus the antithesis of where current research in the field should be going. Quality assurance – with or without the use of algorithmic processes – presupposes *normative standards*. In Asonov et al.'s case, however, the decision in question (DI or NDI) is merely compared with other decisions of the same type, for which there are also no normative criteria of rectitude. Their approach therefore reduces normative criteria (the *Ought*) to empirical phenomena (the *Is*), which is an instance of Hume's naturalistic fallacy. A decision cannot be used as a ground truth against which other decisions can be assessed. A decision can be (un)justified, (un)reasonable etc., but it cannot be true or false. Ultimately, this lack of normative anchor obviously undermines the relevance of subsequent analyses, such as the effect of controlling for environmental factors, some of which Asonov et al. [11] explore, on aggregate measures of model performance. This so-called 'turning the knobs of the model', typical of mainstream ML, does not advance our fundamental understanding of the phenomenon of deception.

The deployment of algorithmic approaches in the high-stake environments of criminal justice and employment practice must be based on defensible multidisciplinary research. The processing of mere convenience data by ad hoc ML models in order to seemingly improve and thus legitimise a scientifically questionable concept such as polygraph-based deception detection does not meet this requirement. Where fundamental rights and principles of justice are at stake, the legal and ethical considerations for scientific research are heightened [1]. The methodological concerns highlighted in this paper must therefore be addressed *before* research can be said to contribute to resolving any of the fundamental validity issues underlying the use of polygraph-based deception detection itself, or any other method to be used in legal proceedings.

CRediT authorship contribution statement

Kyriakos N. Kotsoglou: Writing – review & editing, Writing – original draft. **Alex Biedermann:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

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

Acknowledgments

Funding: Alex Biedermann gratefully acknowledges the support of the Swiss Benevolent Society of New York.

References

[1] K.N. Kotsoglou, M. Oswald, Not "Very English" – on the use of the polygraph by the penal system in England and Wales, J. Crim. Law 85 (3) (2021) 189–208.
[2] P.L. Travers, Mary Poppins Comes Back, in: Mary Poppins - The Complete Collection, Harper Collins Publishers, London, 2008.
[3] K.N. Kotsoglou, Zombie forensics: the use of the polygraph and the integrity of the criminal justice system in England and Wales, Int. J. Evid. Proof 25 (1) (2021) 16–35.
[4] H. Swofford, C. Champod, Implementation of algorithms in pattern & impression evidence: a responsible and practical roadmap, Forensic Sci. Int.: Synergy 3 (2021) 100142.
[5] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.H. Berger, N.W. Brand, R.D. Stoel, Validity and reliability of forensic firearm examiners, Forensic Sci. Int. 307 (2020) 110112.
[6] C. Aitken, F. Taroni, S. Bozza, Statistics and the Evaluation of Evidence for Forensic Scientists, third ed., John Wiley & Sons, Chichester, 2020.
[7] J.S. Buckleton, J.A. Bright, S. Gittelson, T.R. Moretti, A.J. Onorato, F.R. Bieber, B. Budowle, D. Taylor, The probabilistic genotyping software STRmix: utility and evidence for its validity, J. Forensic Sci. 64 (2019) 393–405.
[8] C.T. Bergstrom, D.J. West, Calling Bullshit, The Art of Skepticism in a Data-Driven World, Random House, New York, 2020.
[9] K. Abbasi, Retract or be damned: a dangerous moment for science and the public, BMJ 381 (2023) 1424.
[10] A. Barnett, J. Byrne, Retract or be damned: the "bystander effect" is worsening the situation, BMJ 382 (2023) 1654.
[11] D. Asonov, M. Krylov, V. Omelyusik, A. Ryabikina, E. Litvinov, M. Mitrofanov, M. Mikhailov, A. Efimov, Building a second-opinion tool for classical polygraph, Nature Scientific Reports 13 (2023) 5522.
[12] 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.
[13] A. Biedermann, K.N. Kotsoglou, The unbearable lightness of ignoring axiomatic principles – a response to: "On coping in a non-binary world: rejoinder to Biedermann and Kotsoglou", (by Nicholas Scurich and Richard S. John, in: Statistics and Public Policy, 2024), 2024. https://serval.unil.ch/resource/serval:BIB_456ADCAA77BE.P002/REF.
[14] E.H. Meijer, B. Verschuere, M. Gamer, H. Merckelbach, G. Ben-Shakhar, Deception detection with behavioral, autonomic, and neural measures: conceptual and methodological considerations that warrant modesty, Psychophysiology 53 (2016) 593–604.
[15] British Psychological Society (BPS), Report of the working group on the use of the polygraph in criminal investigation and personnel screening, Bull. Br. Psychol. Soc. 39 (1986) 81–94.
[16] British Psychological Society (BPS), A review of the current scientific status and fields of application of polygraphic deception detection. Final Report from the BPS Working Party Leicester, 2004.
[17] Congress of the United States, Office of Technology Assessment, Scientific Validity of Polygraph Testing: A Research Review and Evaluation, A Technical Memorandum, OTA-TM-H-15, Washington, DC, 1983.
[18] National Research Council (NRC), The polygraph and lie detection. Committee to Review the Scientific Evidence on the Polygraph, Division of Behavioral and Social Sciences and Education, The National Academies Press, Washington, DC, 2003.
[19] K. Alder, The Lie Detectors, Free Press, New York, 2007.
[20] E.E. Jones, H. Sigall, The bogus pipeline: a new paradigm for measuring affect and attitude, Psychol. Bull. 76 (1971) 349–364.
[21] P. Roberts, A. Zuckerman, Criminal Evidence, third ed., Oxford University Press, Oxford, 2022.
[22] K.N. Kotsoglou, Proof beyond a context-relevant doubt. A structural analysis of the standard of proof in criminal adjudication, Artif. Intell. Law 28 (2020) 111–133.
[23] President's Council of Advisors on Science and Technology (PCAST), Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President, Washington, D.C., 2016.
[24] 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.
[25] R. Haraksim, D. Ramos, D. Meuwly, C.E.H. Berger, Measuring coherence of computer-assisted likelihood ratio methods, Forensic Sci. Int. 249 (2015) 123–132.
[26] 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.
[27] A. Biedermann, Machine learning enthusiasts should stick to the facts, response to Morrison et al, Forensic Sci. Int.: Synergy 4 (2022) 100229, 2022.
[28] L. Floridi, J. Cowls, A unified framework of five principles for AI in society, Harvard Data, Sci. Rev. 1 (1) (2019) 1–15.
[29] J. Pearl, The limitations of opaque learning machines, in: John Brockman (Ed.), Possible Minds: 25 Ways of Looking at AI, Penguin Press, New York, 2019, pp. 13–19.
[30] J. Pearl, D. Mackenzie, The Book of Why, Allen Lane, London, 2018.
[31] W. Wundt, Grundzüge der physiologischen Psychologie, fifth ed., Vol. 1, Verlag von Wilhelm Engelmann, Leipzig, 1902.
[32] G.H. Gudjonsson. The Psychology of Interrogations and Confessions, Wiley, Chichester, 2003.
[33] C.D. Lee, The Instrumental Detection of Deception, Thomas, Springfield, IL, 1952.

[34] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning, from Theory to Algorithms, Cambridge University Press, Cambridge, 2014.
[35] H.R. Arkes, J.J. Koehler, Inconclusives and error rates in forensic science: a signal detection theory approach, Law Probab. Risk 20 (2021) 153–168.
[36] A. Biedermann, K.N. Kotsoglou, Commentary on "Three-Way ROCs for Forensic Decision Making" by Nicholas Scurich and Richard S. John (in: Statistics and Public Policy), Statistics and Public Policy 11 (2024) 1–2.
[37] N. Scurich, R.S. John, On coping in a non-binary world: rejoinder to Biedermann and Kotsoglou, Statistics and Public Policy 11 (2024) 1–2.
[38] W.C. Thompson, F. Taroni, C. Aitken, How the probability of a false positive affects the value of DNA evidence, J. Forensic Sci. 48 (2003) 47–54.
[39] D.A. Schum, Evidential Foundations of Probabilistic Reasoning, John Wiley & Sons, New York, 1994.
[40] I.E. Dror, N. Scurich, (Mis)use of scientific measurements in forensic science, Forensic Sci. Int.: Synergy 2 (2020) 333–338.
[41] F. Taroni, A. Biedermann, P. Garbolino, C. Aitken, A general approach to Bayesian networks for the interpretation of evidence, Forensic Sci. Int. 139 (2004) 5–16.

Kyriakos N. Kotsoglou
University of Northumbria, School of Law, Newcastle Upon Tyne, NE1 8ST, UK

Alex Biedermann *
University of Lausanne, Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, 1015, Lausanne–Dorigny, Switzerland

* Corresponding author.
E-mail address: alex.biedermann@unil.ch (A. Biedermann).