Science & Justice 59 (2019) 1–8

Contents lists available at ScienceDirect

Science & Justice

journal homepage: www.elsevier.com/locate/scijus

Resolving differing expert opinions

Isabelle Montani*, Raymond Marquis, Nicole Egli Anthonioz, Christophe Champod

School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, Batochime, quartier Sorge, 1015 Lausanne-Dorigny, Switzerland

ABSTRACT

This paper explores procedural mechanisms to resolve differing conclusions when two experts have initially worked independently. These experts can be two human examiners or one of them may be a computer-based model. The resolving process is presented as part of the ACE-V protocol adopted widely in pattern recognition areas (e.g. fingerprints, footwear marks, toolmarks or handwritings/signatures comparisons). It set the conditions of operations and delineates a resolving process that is based on the principles of transparency and detailed argumentations. We predict a gradual but steady introduction of computer-based models in the forensic pattern recognition areas. In our opinion, the rules to resolve differing opinions ought to be articulated and documented in the form of standard operating procedures, before any deployment in casework practice.

1. Introduction

In forensic pattern comparison areas, such as fingerprints, toolmarks, footwear marks or handwritings and signatures comparisons, the examination follows a four-step process named ACE-V for Analysis, Comparison, Evaluation and Verification. First formalised by Huber [18], several authors further refined and detailed this process [3,6,34,35]. Some even scrutinized it through experimental research (e.g. [20]). In the first part of this paper, this framework will be briefly introduced, focusing on the Verification step. In spite of existing definitions of this step, advice is scarce regarding the way differing opinions between forensic examiners in the verification stage can be resolved.

Indeed, every expert has faced at one time or another of his career this situation: one expert reaches a different conclusion than the one reached by his colleague. For example, in a pattern comparison case (involving for example toolmarks, footwear marks or fingermarks), a first expert may reach an inconclusive conclusion, while the other may support the proposition of common source. The difference between the two conclusions can be even more marked, for example, if one expert is more inclined towards a common source while the other is guiding towards different sources. A few case examples of differing conclusions between two experts are illustrated in Fig. 1. All throughout this paper, the strength to be attached to the forensic findings – named generically *strength of the forensic findings* or SFF) – is expressed by a likelihood ratio (LR), which is now generally accepted among forensic scientists as a rigorous and logical framework [1,4,14,33,36]. The likelihood ratio is the metric that, when expressed in log10 form, gives the weight of evidence (WoE) to be associated to the forensic observations [15]. This approach clearly identifies the roles and responsibilities of the forensic examiner and of the Court. In this approach, the findings are evaluated in the light of competing propositions that generally reflect the views of the parties at trial [9].

The key question that we explore in this paper is how are these cases with differing conclusions between the two experts resolved? One of the two experts can try to convince the other that his logic is better, he could invoke experience, or that he spent more time on the case, and so on. But if we restrict to a scenario where the two experts have similar training and experience, have spent the same amount of time on the case, and are more or less equal in other aspects relevant to the examination process, how can their differing conclusions be resolved?

Furthermore, the development of automated tools in the forensic pattern evidence examination requires a formalization of the examination process (including the verification) in order to be able to seamlessly integrate these technological developments. Indeed, in most areas of forensic science, automated or semi-automated computer-based tools have been developed to support experts during the examination of pattern evidence. In the near future, in fields essentially based on human judgement, an increased development of computer-based techniques is expected leading to a higher level of distributed cognition between machine and human [12].

That increase is not only a consequence of improved machine learning techniques but also a response to the call for more user-independent techniques in the forensic pattern comparison areas (e.g. [29,30]).

According to Atanasiu [5] “computing is expected to play a central role in shaping the expert opinion” (Atanasiu, pp.75–76). Recent computing developments applied to forensic science indeed concentrated on the evaluation process. Presently, knowledge-based

* Corresponding author.
E-mail address: isabelle.montani@alumnil.unil.ch (I. Montani).

https://doi.org/10.1016/j.scijus.2018.10.003
Received 29 March 2018; Received in revised form 4 October 2018; Accepted 8 October 2018
1355-0306/ © 2018 The Chartered Society of Forensic Sciences. Published by Elsevier B.V. All rights reserved.

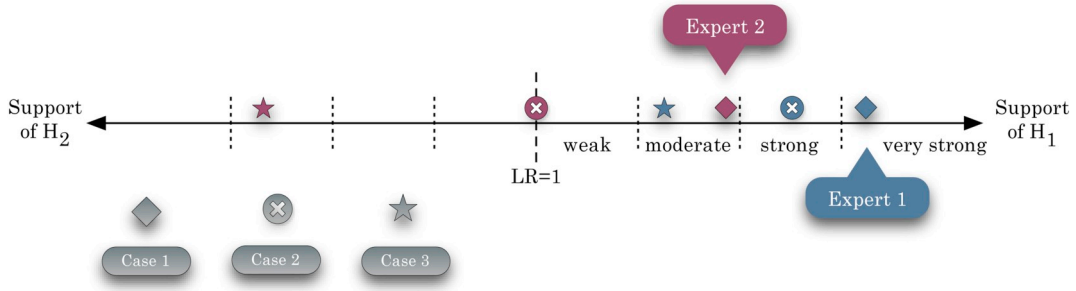

Fig. 1. Differing conclusions reached by two experts (expert 1 and expert 2) in three cases. In Case 1 (diamonds), both experts point towards the same proposition but with different strength. In case 2 (crosses), the first expert leans strongly towards H_1 as opposed to H_2, whereas the second expert does not guide in any direction. In Case 3 (stars), the first expert indicates that the observations would support H_1 as opposed to H_2, the second expert does just the opposite.

systems already operate in the evaluation phase of forensic examination, especially in the field of glass and DNA evidence [10]. Interesting developments have also been published in the area of fingerprints (e.g. [13,26,27]), firearms (e.g. [31]) and handwriting evidence (e.g. [11,16,22,25]).

While the expert of tomorrow will undoubtedly be assisted with technological systems that will help appreciate the strength of forensic findings, little effort has been made to enable forensic examiners to combine model outputs with opinions resulting from a traditional approach in a coherent and rigorous manner. If such a model is developed to be ubiquitously used, then the expert must possess a logical and documented strategy to integrate the model's findings into his own reasoning process. Indeed, without a clear structure clarifying how to embed the two examinations, the addition of each taken individually is worth less for the expert than the two put in combination with each other. Thus, in the third part of this paper we will try to address how to link both the traditional holistic examination expertise and a developed quantitative model.

A framework will be proposed and may be used as a general guideline to help handle conflicting scenarios between two experts or between an expert and a model. This will also be useful for forensic examiners to transparently structure and communicate their views. Examples will be proposed and discussed in forensic fields where human visual perception takes an important part in the ACE-V approach, i.e. for pattern evidence, with a focus on two disciplines: handwriting and fingerprint evidence.

It is important to stress the difference between the use of computed-based assessment techniques in a field such as DNA profiling compared to pattern recognition areas such as handwriting or fingerprints. Models developed to assign WoE to DNA findings are not used in conjunction with an appraisal by the expert of that weight. As soon as the model underpinning the calculation is accepted as validated for casework, the value obtained from the model will be the value quoted in the DNA statement. There is no competition between a holistic assessment of a DNA expert and the model. In pattern comparison areas however, the assignment of the WoE has been entirely left for years to the expert's judgement without any computer-based support. Experts developed specific recognition expertise based on a large set of features [32]. The computer-based models developed in these areas are largely based on a limited (but tractable) subset of features. It means that the computed WoE cannot stand alone as the only guidance towards the overall weight but will have to work in concert with the holistic judgement of the expert.

2. ACE-V Stepwise approach and its use in pattern evidence evaluation

The forensic examination process of a pattern evidence is divided into three steps, ACE, that are the analysis, the comparison and the evaluation process, as presented by Huber and Headrick [19], based on the early publications by Huber [17,18]. The concept of conducting a sequential set of tasks distinguishing analysis from comparison goes back to the early days of forensic science [20]. The verification (V) step was subsequently added by Ashbaugh [2] for fingerprint examination and adopted on most pattern comparison areas. The final stage is carried out by a second expert.

This whole process (schematically presented in Fig. 2) is now commonly referred to as ACE-V and has become a widespread approach applicable in a general manner to all types of forensic pattern evidence.

The ACE-V is not a method per se, but a stepwise approach to guide examiners throughout their examination process [7]. The expert thus independently analyses the questioned and the reference material, and afterwards compares the results of the findings of both entities. The features that are retained at the end of the comparison stage for the evaluation stage are considered as the forensic findings (FF). The forensic findings are evaluated in light of the given propositions, and assigned a strength that will allow the expert to support one of the two propositions with respect to the other.

The second expert often carries out only the final stage of the examination process, the verification phase, by critically looking at the findings of his colleague and reading the proposed report, followed by a thorough joint discussion of their conclusions (see Fig. 3).

This way of proceeding is time efficient and offers the benefits of a peer-review (adoption of a four-eyes principle), but it suffers from an obvious risk of confirmation bias (see Cole [8]). As the risk of confirmation bias needs to be mitigated, we would require both experts to conduct two independent ACE processes (see Fig. 4). The whole examination procedure should then be carried out by both experts, thus guaranteeing independence between the outcomes with a realistic and meaningful confrontation of both experts' results. In this case, both experts carry out the ACE procedure, either one after the other, or simultaneously, and their respective results are confronted in the V stage.

Fig. 2. The ACE-V pattern evidence examination procedure comprising the analysis of the questioned and reference material, their comparison, the evaluation and verification stages.

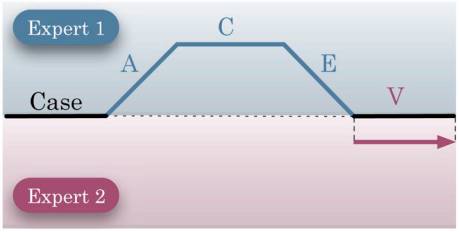

Fig. 3. Schematic representation of an ACE examination procedure. The first expert carries out the ACE stages, and the second expert only carries out the V stage (based on the findings and conclusions of the first expert).

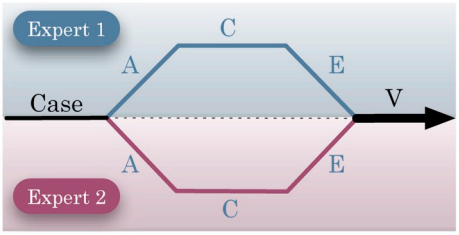

Fig. 4. Schematic representation of the ACE-V examination procedure. Both experts carry out the ACE stages, and confront their results in the V stage.

The simultaneous two-expert ACE-V examination has the advantage of minimizing the potential bias effects since the second expert will not have the temptation of blindly adhering to the first experts' results.

However, if such a procedure is carried out, the evaluation and verification stages must be transparently dissected in order to coherently propose a scheme for the joint conclusion, especially when both experts would come to differing conclusions at first. An evaluation process enabling two experts to arrive at a consensual final conclusion, and taking into account the specifics of both the evaluation and verification stages, is therefore essential. The different analyses, confrontations and discussion steps that should take place between two (or more) experts in their evaluation procedure should be deconstructed, allowing for their logical arrangement in a transparent manner.

3. Process to reconcile differing conclusions between two experts

The proposed working procedure is established in the context of a best-case scenario, where one expert has been mandated to work on a case, and does so with the ACE-V approach. The second expert is equivalent in terms of training and experience, uses the same material, and both of the experts are expected to give a final consensual conclusion on the pattern evidence case.

For this general canvas, and in direct relation to each specific case, the propositions are known and are identical for both experts working the case. Both experts must carry out the evaluation according to the same underpinning probabilistic principle, and must provide a conclusion on a common scale.

As stated beforehand, the propositions for each case are defined; these propositions will be identical for both experts working the case. Following the analysis and comparison stages, a set of forensic findings (FF) is available. We propose a verification procedure where the forensic findings (FF_1 and FF_2 in Fig. 5) may differ between experts; this is an approach to verification where two experts carry out the whole ACE process. However, the schematic representation in Fig. 5 is just as applicable if expert 2 only attributes strength to the forensic findings defined by expert 1; in this case expert 2 starts the evaluation process with FF_1. The evaluation by experts 1 and 2 gives rise to the resulting strengths of the evidence, SFF_1 and SFF_2, conditioned by the propositions that are defined by the case. At this point, the two elements SFF_1 and SFF_2 are confronted: does the strength both experts have attributed to the forensic findings indicate a substantially identical outcome? A

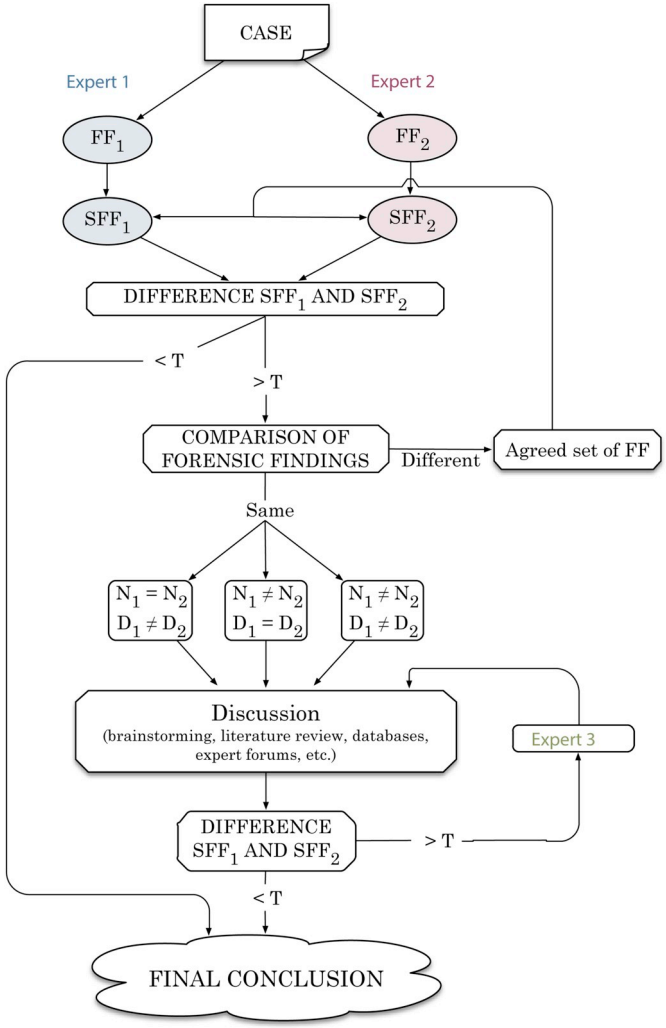

Fig. 5. Schematic representation of the integration of two different experts working on a same case.

threshold (T) needs to be defined in order to evaluate differences between SFF_1 and SFF_2. If this difference is below the set threshold, the final conclusion can be formulated directly given this strength of forensic findings. It is important to note that this threshold is defined beforehand, for example as a laboratory policy, and not adjusted on a case-by-case basis. Also, and given that the two assessments are substantially identical, the responsibility of formulating the definitive conclusion falls to the expert named as such in the case. Setting the threshold T is a matter of policy and assessment as to the impact of a variation of SFF on the decision maker. When likelihood ratios are expressed in log10 terms (as WoE), a value of ± 1 represents, in our opinion, an adequate option. It is customary, when likelihood ratios are mapping into verbal equivalent to adopt a logarithmic scale [23]. A difference on SFF that would impact the choice of verbal equivalents is a good way to define T.

If the difference between SFF_1 and SFF_2 exceeds the threshold, in a first step, the forensic findings FF_1 and FF_2 are compared to each other. If these forensic findings are different, the two experts need to reach an agreement on the common set of forensic findings, and return to the evaluation stage, yielding a new assessment of the strength of the findings using this new, agreed, set of forensic findings.

If the strength of forensic findings is substantially different while the forensic findings are essentially the same, the difference in the assignment of value to these findings needs to be tracked through the different

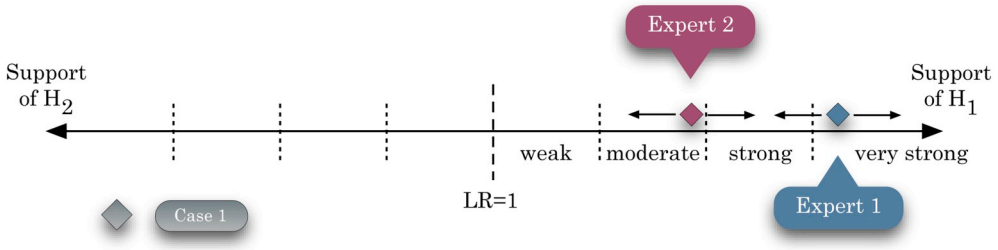

Fig. 6. Illustration of how two experts facing different conclusions can adapt their conclusions.

elements allowing this assignment.[1] The evaluation of forensic findings consists in the assessment of the probability of these findings under two competing propositions. One is the probability of the forensic findings given the first proposition (or numerator of the LR, N_1 and N_2 in Fig. 5) [33]. The second is the probability of these same findings given a second proposition (or denominator of the LR, D_1 and D_2 in Fig. 5). The difference in the assessment of the overall strength of the evidence can be due to a difference in either the numerator or the denominator of the LR, or in both. The source of the difference in the numerator ($N_1 \neq N_2$) between both experts can reside in their assessment of the variation in the reference material or in their assessment of the importance of a perceived dissimilarity between the questioned and reference material. A difference in the denominator, $D_1 \neq D_2$ on the other hand, can be due to the respective assessments of the rarity of the forensic findings. Once it is known whether the numerator, the denominator, or both are at issue, the difference shall be meaningfully discussed, in the form of a brainstorming activity, by researching and consulting literature, conference proceedings, articles and books and/or by consulting databases or expert forums or crowd-source expert opinions (for example specific working groups). Each of the expert's evaluations (and therefore conclusions) can be moved along the scale of the support of the proposition (commonly referred to as hypothesis H_1 or H_2). The global aim of finding a reasoned, consensual conclusion between both experts is to have the experts slide along the scale (upwards or downwards such as presented in Fig. 6) to a point where they are both content and satisfied with the final conclusions.

At the end of this process, which is not meant to be a simple validation of the opinion of one expert by the other but a scientific procedure based on data and reasoning, the difference in the strength of forensic findings assigned by the two experts is either smaller than the threshold, leading to the conclusion of the process, or larger. At this point, it may be an option to report both differing conclusions of the two experts, accompanied with motivated arguments. However, in agreement with Fig. 5, we strongly recommend the intervention of a third expert, who may be an individual or a pool of experts, such as an expert committee. This additional expert adds her opinion to the discussion on the relative probability of the forensic findings under both propositions evaluated. Working from the evaluation phase, the third expert does not have to redo the whole process from the beginning, seeing as both experts 1 and 2 have already reached a consensus regarding the forensic findings (FF). The third expert must however be involved in the assessment of the numerator and denominator of the LR.

The discussion with the third expert should allow the assignment of the strength of the forensic findings with a difference smaller than the threshold for the first two experts. However, should this not be the case, no final consensual conclusion can be given. The experts may then state their different opinions in the report, and motivate their diverging conclusions. The final conclusion (or set of conclusions) is thus transparent for the receiving audience. In this case, it is not appropriate for the named expert to overrule divergent opinions and report only his own assignment of the strength of the evidence.

3.1. Applied example – signature

This example is taken from the 14th ENFHEX Collaborative Exercise. Experts were requested to give opinions regarding six questioned signatures, on the basis of comparison with ten reference specimens. Experts were provided with scans of both the questioned and the reference signatures.

Let us consider a disagreement between conclusions of two experts, concerning a given questioned signature. According to the first expert, the results strongly support (LR between 100 and 1000) the proposition that the questioned signature is genuine rather than a forgery.[2] In the view of the second expert, the results only moderately support (LR between 10 and 100) the proposition that the questioned signature is genuine compared to a forgery.

Scrutiny of the findings of both experts revealed that they formed their opinion on the basis of the same forensic findings. Their disagreement can be explained by a different evaluation of these forensic findings. The second main part of the questioned signature presents an obvious pen lift, which is never observed on the reference signatures (see Fig. 7). In the opinion of the first expert, the complexity of the signature is high and the combination of all similarities is not likely under the proposition of a forgery, despite the presence of the pen lift. From the point of view of the second expert, given the moderate complexity of the signature, the presence of this pen lift, altogether with the similarities observed, are quite probable if the questioned signature is a forgery. Hence, the two experts disagree on the magnitude of the denominator of the likelihood ratio.

After discussion, once both experts realised the source of their divergence, they decided that not too much weight must be given to the pen lift (in respect with all similarities), and that the many direction changes and crossing points confer a certain complexity to this signature. Both experts finally agreed on the conclusion that the results strongly support the proposition that the questioned signature is genuine as opposed to forged. The case file will reflect the whole process. The initial conclusions reached by both experts should be available alongside with the minutes of the discussion and arguments used to jointly decide on the conclusion that will be reported in this case. The legitimacy of the whole process is critically dependant on the extent and thoroughness of this documentation.

The results of the collaborative exercise revealed that this questioned signature was indeed genuine.

3.2. Applied example – fingerprint

For this example, examiners 1 and 2 both examine the fingermark and fingerprint below. The conclusions offered by the experts are, respectively, that the elements observed very strongly support (SFF_1

[1] Please note here the importance of case notes indicating the thought process leading to the assessment of the strength of the forensic findings in a transparent manner.

[2] In all case examples, we will refer to the verbal scale presented in Marquis et al. [23].

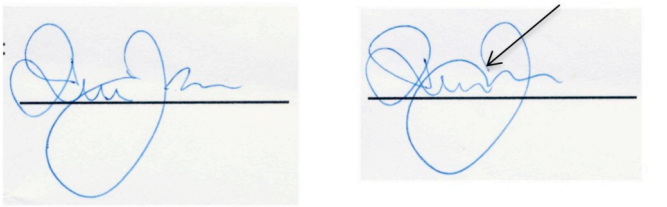

Fig. 7. Signature case; a pen lift is visible in the second part of the questioned signature (see arrow), which is absent from the reference material (only one reference is shown here).

amounting to a LR between 1000 and 10′000) the proposition of a common source versus different sources and that a proposition of common source is excluded (SFF_2 amounting to a LR of 0). Cleary the difference is larger than any reasonable threshold T.

While overall the features used by the experts (compared after the difference in assignment of strength has been detected) are the same, the difference lies in the assignment of the numerator probability. While examiner 1 assigns a medium probability that the findings (similarities, dissimilarities) would be observed if the mark and the print come from the same source, the second examiner assigns an extremely low value to this probability. For examiner 2 this is impossible, amounting to a probability of 0. The features that are mainly responsible for this different assessment are shown in Fig. 8. After seeing the disagreement illustrated, where features are absent from the fingerprint that are present on the mark, examiner 2 has stopped his assessment, judging that the assignment of a denominator wasn't relevant anymore. A discussion of the assignment of the numerator has been conducted between the two examiners. For examiner 1, the difference is due to an artefact created by the inking process, leading to continuous ridges above the center, while the mark is a better representation of reality. For examiner 2, these differences are unexplained. In his opinion, there is no indication on the inked print to show any problem; therefore, the numerator is so low as to exclude a common origin. During the discussion, the two examiners agreed on checking on another impression of the same finger; on this new inked impression, the arrangement of minutiae in the center is compatible with that observed on the mark; therefore, the difference was indeed due to the inking process. Examiner 2 goes back to the assignment of a denominator value, and comes to the same strength assignment as examiner 1, that is to say, very strong support for a same source between mark and print as opposed to different sources.

4. Process to reconcile differing conclusions between an expert and a computer-based model

Resolving conflicting conclusions between an expert and a model is an extension of the proposal made between two experts. In fact, we could assimilate the computed-based model to a dedicated expert who carried out its own evaluation. It is assumed here that the propositions against which the forensic findings are evaluated are the same for both the human examiner and the model. A further premise to the approach proposed here is that the features considered by the model are of the same type as those used by the expert; that is to say, the model is not a black box, but rather, the features used are human-readable, and of the same nature as those used by the examiner. It means that there should be a possibility to assess if the features used by the human expert differ or not from the features used by the computer-based model.

The whole process is presented in Fig. 9. The forensic findings, together in regard to the propositions from the case, lead both the expert and the model to an evaluation of the strength of the forensic findings given these propositions. We call them SFF_1 and SFF_M for the expert and the model, respectively. The next step (whether the evaluations correspond substantially or not) is to assess whether the examiner considered all the features integrated into the model. If not, she goes back to the features and carries out a new evaluation, leading to an updated strength of the forensic findings.

We acknowledge the fact that, in pattern comparison areas, the features available to the examiner exceed, at this time, those used by the models. Therefore, an evaluation of whether, given only the features used by the model, SFF_1 and SFF_M correspond (that is, their difference is below a previously set threshold T as previously suggested) is carried out. If this is not the case, the examiner shall adopt the models' result. Saying that, all features considered equal, the output of the model shall be preferred to the holistic judgement of the human expert is a strong claim that requires some explanations. It goes at the hearth of computer-based models and their validation in forensic practice.

There are two critical conditions that need to be met to develop a meaningful resolution process. The first relates to the recognition given to the computer based model to provide meaningful guidance. In our discussion regarding the case where two human-experts would confront their results, we specified that both experts would be considered equivalent in terms of training and experience, hence in their ability to derive conclusions that would be trusted equally in the resolution process. The same should apply with a computer-based model. The model should have been signed-off to be use in casework through a proper validation process. By validation we mean that the model should have been assessed empirically against ground truth data and appropriately calibrated. Typically, it would be expected for it to meet the requirements set in the guideline proposed by Meuwly et al. [24]. Part

Fig. 8. Fingerprint case; characteristics leading to different assignments in the numerator are indicated with arrows. On the far right, a new reference fingerprint showing the same characteristics as the mark, resolving any ambiguities.

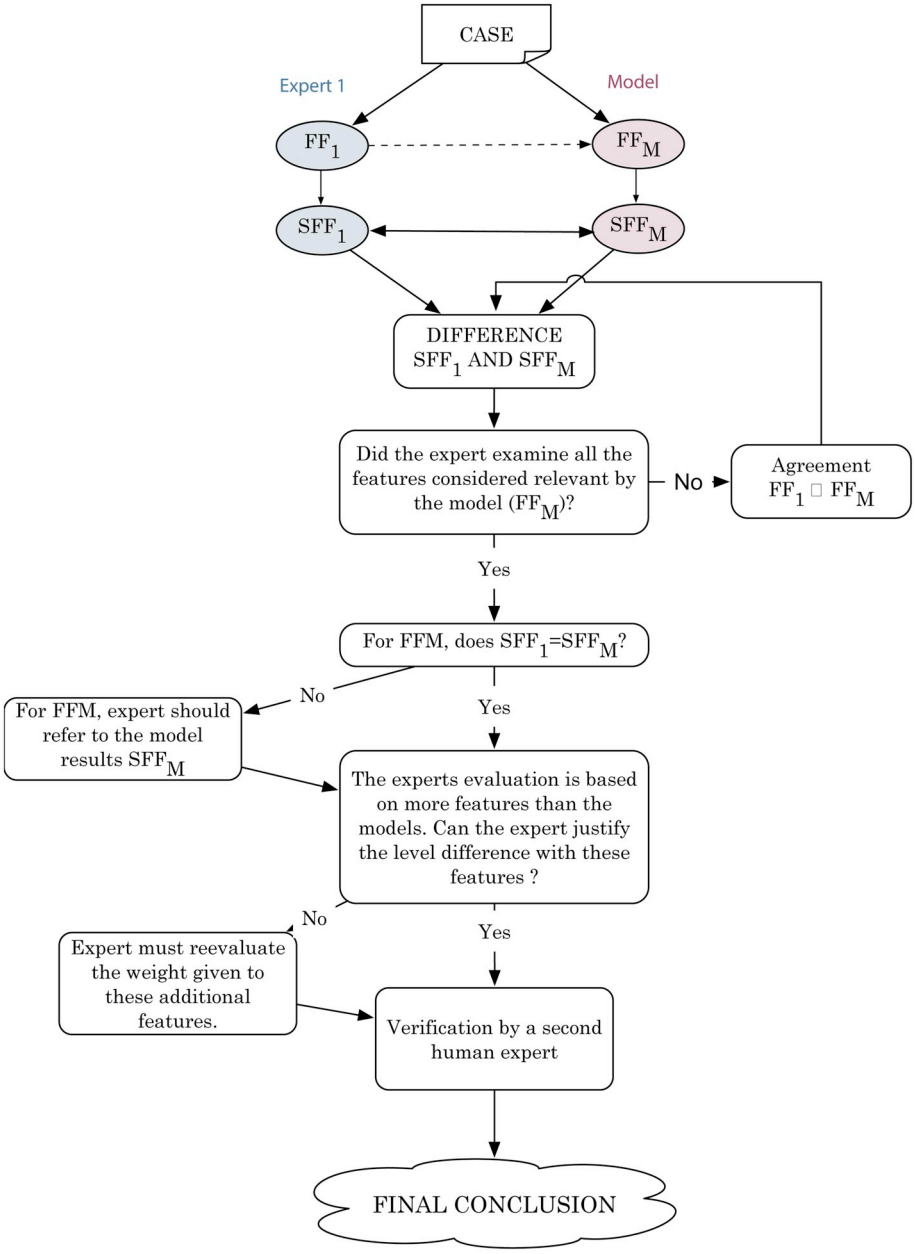

Fig. 9. Schematic representation of the integration of the results of a human expert and a model. It is often the case that the expert examines more features than the model. However, if the expert and model features are exactly the same, the expert should simply refer to the SFF provided by the model. Such a scenario is not represented.

of the validation will also aim at defining the boundaries of usage, recognizing that a given model may not apply to all cases. It means that prior using a model in a case, and reaching the verification stage discussed here, there is a prior decision that agreed to use the model in this case. Once that decision is made, the output of the model cannot be dismissed on the ground, for example, that the model was not adapted for this case. The second requires a common assessment scale. Indeed, both experts must carry out their evaluation according to the same underpinning probabilistic principle, and must provide a conclusion on a common scale. Throughout this paper we gave precedence to a likelihood ratio based approach.

If both of the above-conditions are met and that for a given set of features (FF_M), both assessments SFF_1 and SFF_M are divergent, which assessment should be given greater consideration? Our position is that it should be the model. The final assignment might be adjusted, as we shall see, in a later stage based on the observations of additional features, but for the shared set of features considered (FF_M), the model output shall prevail. SFF_1 has been informed by a holistic assessment based on training and experience, for example accounting for the rarity in the relevant population of a given set of features. The model offers a systematic measure based on documented and user-independent corpus that has been validated for casework. The model output has better scientific credential and it should calibrate the human judgement. Systematic data when available, acquired under controlled conditions, shall always take precedence over experiential based judgements.

In a further resolution step though, the examiner may adapt the models' results if she uses additional characteristics, which are not considered by the model. This, of course, requires from the expert knowledge of within- and between- source variability of the additional features considered by the examiner. To restate the points made previously, that flexibility offered to the expert is not a licence to move up or down the SFF without any accountability. If the examiner chooses to

adapt the strength based on additional features, they should be documented and the reasoning and underpinning data provided in a transparent way. She may invoke systematic data (not captured by the model), specific documented knowledge or personal experience. In any case, the basis of the adjustment of the strength to be attached to the findings should be clearly stated. Referring back to the ACE-V process, that latter stage would require an independent verification by another examiner.

If there are no additional features, or if the additional features used by the examiner do not warrant a difference in the strength to be assigned to the findings, it will remain as computed by the model.

4.1. Applied example – fingerprint

A mark of questioned origin shows 7 minutiae in correspondence with a given print. These 7 minutiae are not particularly rare (ridge endings flowing out from the delta). This, for the fingerprint examiner, leads her to conclude to a moderate support of a common source for mark and print (as opposed to different sources). The model outputs a likelihood ratio of 4000. This falls within compatible orders of magnitude. Examiner and model have considered the same features. However, in addition to these 7 minutiae, the examiner noted (already in the analysis stage) a scar that is visible on the mark and can be found on the print. At this stage of the computer-based model development, this scar is not integrated. The model is considering only the strength associated with the minutiae. The consideration of the scar increases the strength of the forensic findings, given that its shape, beginning and ending, the exact path followed across the ridges are clearly visible on both impressions. Therefore, the examiner assesses the overall likelihood ratio as higher than the output from the model, leading to strong support for a common source rather than different sources. The factor of 10 given by the examiner to the scar correspondence is based on training and experience as, to her knowledge, the variability between scars on different fingers has not been systematically studied. That final assessment has been signed-off by a second examiner following an independent examination (and, if needed, the first resolution process previously detailed).

4.2. Applied example – signature

In the signature case presented in the previous section of this paper, let's assume a model-based approach was used in the place of the second expert. Examiner 1 concluded to ‘strong support’ for a genuine signature, while the model output a likelihood ratio of 20′000 (‘very strong support’). The expert, upon comparing his own observed features to those used by the model, observes that his features cover all of the features used by the latter. These features are the proportions between the letters and the angles of the line strokes. According to his notes, the expert does not arrive at the same strength as the model based on FF_M. Further analysis of this disagreement shows that his numerator (probability of the observed features if the signature is genuine) is similar to that output by the model. The difference in the LR is due to his assessment of the denominator, expressing his belief in the complexity of the signature, the difficulty to imitate it, which is a larger probability than that output by the model. Given that the model output is based on a large database allowing an assessment of this denominator is based on solid structured data, the expert will adjust his own assessment of the denominator accordingly.

The expert retains the LR value from the model for the features it uses (FF_M) but additionally considers features not integrated into the model (overall shape, pressure, formation). In this case, the overall conclusion remains ‘very strong support’, since the expert did not feel that the additional contribution of these features will significantly impact the strength of the forensic findings.

5. Discussion and conclusion

Through decomposition of the examination process of forensic scientists in order to facilitate the identification of any disagreement source, the authors have contributed to clarify what the V of the ACE-V process stands for. It was shown that the verification step does not simply represent a confirmation that validates the view of an expert. As stated by Ashbaugh [3], verification is rather “a form of peer review”, which involves the critical inspection of the processes and reasonings of the experts [20]. In this paper, we presented a verification stage of the forensic examination process as being operated through intervention of a second examiner or a model.

The resolution protocols offer general guidelines for harmonizing the work between experts or between an expert and a model. The proposed resolution process is no different from what forensic laboratories have put in place. However, we have often witnessed that the conclusion bearing the less weight in favour of the prosecution's view is retained as the reported conclusion on the grounds of being conservative and erring on the safe side. Our view is that the consensus shall be based on sound scientific arguments and not on policy. It means that if no consensus is obtained, even after consultation with a third expert, the case will be reported with an indication of all opinions. It may be seem odd to report differing conclusions in the same case, but that level of transparency is required by the complexity of the case at hand.

The protocols presented in this paper are to be understood as a framework for integrating results of different sources (whether the conclusions come from experts or from an expert and a model). The transparency required by the process allows for a juxtaposition of both model and expert conclusions. With such a transparent system, any decision of a different conclusion than the one given by the model or a first expert must, and more importantly can, be justified by the expert.

Both protocols developed help the expert better understand the parameters that have an impact on his conclusions and on the conclusions of his colleagues. The developed protocols thus offer coherent frameworks for forensic scientists to properly handle the complexity of differing opinions, and substantially help structure and communicate their views in the most transparent manner possible.

Dealing with the situation involving an expert and a model, extensive knowledge of the workings of the model, as well as its limitations, are required for proper integration of its results into the forensic process. Furthermore, it must be emphasized that once the procedure of implementing the quantitative model into the traditional examination process is finalized, picking and choosing, for example, cannot be an option. The expert cannot take into account the results of the model only when they provide him with a supplementary weight in his conclusions, comforting him in his earlier decision. Likewise, the expert cannot decide to discard the results of the model that do not please him, or that go against the results obtained with the holistic examination. Therefore, the possible conflicting scenarios must be identified.

Our vision for the future is that the gap between the features considered by the models and the features considered by the human expert will gradually reduce. This will be due to the development in machine learning and forensic modelling efforts. We can foresee the time where the output of the model will take precedence over the human informed opinion in the assignment of the strength to be attached to the forensic findings. But it is still a stepwise approach and, in the meantime, forensic examinations in pattern comparison fields, will remain a cohabitation between human and machine. Hence the need to specify the rules of engagements and procedures to deal with conflict resolution.

Further research should also focus on resolution of differing opinions between an expert and a model based on features that are not human-readable. Typically, in the area of face recognition, facial comparison systems are extracting from images features that are generally not meaningful to a human observer. In these cases, the whole concept of comparisons the retained forensic findings (FF) that we have

suggested here, will become moot. Finally, in this paper we have dealt with the joint work of a human expert with one computer-based model. In the future, a range of models may be available and of equal scientific merit. It will lead to potential differing results that will need to be resolved as well.

References

[1] C. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., John Wiley & Sons, Ltd, Chichester, UK, 2004.
[2] D. Ashbaugh, Ridgeology, J. Forensic Identif. 41 (1991) 16–64.
[3] D. Ashbaugh, Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology, CRC Press, Boca Raton, FL, USA, 1999.
[4] Association of Forensic Science Providers, Standards for the formulation of evaluative forensic science expert opinion, Sci. Justice 49 (2009) 161–164.
[5] V. Atanasiu, Expert Bytes: Computer Expertise in Forensic Documents – Players, Needs, Resources, and Pitfalls, CRC Press, London, New York, 2014.
[6] M. Cassidy, Footwear Identification. Royal Canadian Mounted Police, Canadian Government Publishing Centre, Hull, Québec, Canada, 1980.
[7] C. Champod, Fingerprint examination: towards more transparency, Law Probab. Risk 7 (2008) 111–118.
[8] S.A. Cole, Implementing counter-measures against confirmation bias in forensic science, J. Appl. Res. Mem. Cogn. 2 (2013) 61–62.
[9] R. Cook, I. Evett, G. Jackson, P. Jones, J. Lambert, A hierarchy of propositions: deciding which level to address in casework, Sci. Justice 38 (1998) 231–239.
[10] J.M. Curran, Allan Jamieson, André Moenssens (Eds.), Use of Knowledge-Based Systems in Forensic Science, Wiley Encyclopedia of Forensic Science, John Wiley & Sons, Ltd., Chichester, 2009, pp. 2590–2593.
[11] L. Davis, C. Saunders, A. Hepler, J. Buscaglia, Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios, Forensic Sci. Inte. 216 (2012) 146–157.
[12] I. Dror, J. Mnookin, The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensics, Law Probab. Risk 9 (2010) 47–67.
[13] N. Egli, Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System, PhD Thesis School of Criminal Justice, Faculty of Law and Criminal Justice, University of Lausanne, Switzerland, 2009.
[14] I. Evett, Interpretation: a personal odyssey, in: C. Aitken, D. Stoney (Eds.), The Use of Statistics in Forensic Science, Ellis Horwood, Chichester, UK, 1991, pp. 9–22.
[15] I.J. Good, Probability and the Weighing of Evidence, Charles Griffin & Company Ltd., London, UK, 1950.
[16] A. Hepler, C. Saunders, L. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int. 219 (2012) 129–140.
[17] R.A. Huber, Expert witnesses, Crim. Law Q. 2 (1959) 276–295.
[18] R.A. Huber, The philosophy of identification, RCMP Gazette (1972) 9–14.
[19] R.A. Huber, A.M. Headrick, Handwriting Identification Facts and Fundamentals, CRC Press, Boca Raton, USA, 1999.
[20] G. Langenburg, A critical Analysis and Study of the ACE-V Process, PhD Thesis School of Criminal Justice, Faculty of Law and Criminal Justice, University of Lausanne, Switzerland, 2012.
[22] R. Marquis, Etude de caractères manuscrits: de la caractérisation morphologique à l'individualisation du scripteur, PhD Thesis School of Criminal Justice, Faculty of Law and Criminal Justice, University of Lausanne, Switzerland, (2007).
[23] R. Marquis, A. Biedermann, L. Cadola, C. Champod, L. Gueissaz, G. Massonnet, W.D. Mazzella, F. Taroni, T. Hicks, Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings, Sci.Justice 56 (2016) 364–370.
[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] I. Montani, Exploring Transparent Approaches to the Authentication of Signatures on Artwork, PhD Thesis School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, Switzerland, 2015.
[26] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, J. Forensic Sci. 52 (2007) 54–64.
[27] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm, J. Royal Statis. Soc. A 175 (2012) 371–415.
[29] President's Council of Advisors on Science and Technology, Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President, Washington DC, 2016https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
[30] President's Council of Advisors on Science and Technology, An addendum to the PCAST report on forensic science in criminal courts, Executive Office of the President, Washington DC, 2017https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf.
[31] F. Riva, C. Champod, Automatic Comparison and Evaluation of Impressions Left by a Firearm on Fired Cartridge Cases, J. Forensic Sci. 59 (2014) 637–647.
[32] R.A. Searston, J.M. Tangen, Expertise with unfamiliar objects is flexible to changes in task but not changes in class, PLoS ONE 12 (2017) e0178403.
[33] F. Taroni, C. Aitken, P. Garbolino, A. Biedermann, Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science, 2nd edition, John Wiley & Sons, Ltd., Chichester, UK, 2014.
[34] H. Tuthill, Individualisation: Principles and Procedures in Criminalistics, Lightning Powder Company, Salem, OR, USA, 1994.
[35] J. Vanderkolk, ACE + V: A model, J. Forensic Identif. 54 (2004) 45–52.
[36] S. Willis, Kenna L. Mc, Dermott Mc, G. O' Donnell, A. Barrett, B. Rasmusson, T. Höglund, A. Nordgaard, C. Berger, M. Sjerps, J.J.L. Molina, G. Zadora, C. Aitken, T. Lovelock, L. Lunt, C. Champod, A. Biedermann, T. Hicks, F. Taroni, ENFSI Guideline for Evaluative Reporting in Forensic Science - Strengthening the Evaluation of Forensic Results across Europe, http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf, (2015).