Law, Probability and Risk (2010) **9**, 47–67 doi:10.1093/lpr/mgp031
Advance Access publication on January 22, 2010

The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science

ITIEL E. DROR†
Institute of Cognitive Neuroscience, University College London (UCL) and Cognitive Consultants International (CCI) Ltd, 17 Queen Square, London, WC1N 3AR, UK

AND

JENNIFER L. MNOOKIN‡
Professor of Law, UCLA School of Law, 405 Hilgard Ave, Los Angeles, CA 90095, USA

[Received on 14 May 2009; revised on 19 October 2009; accepted on 3 November 2009]

Cognitive technologies have increased in sophistication and use, to the point of interactively collaborating and distributing cognition between technology and humans. The use of Automated Fingerprint Identification Systems (AFIS), computerized databases of fingerprints, by latent fingerprint experts, is a par-excellence illustration of such a partnership in forensic investigations. However, the deployment and use of cognitive technology is not a simple matter. If a technology is going to be used to its maximum potential, we must first understand the implications and consequences of using it and make whatever adaptations are necessary both to the technology and to the way humans work with it. As we demonstrate with AFIS, latent fingerprint identification has been transformed by technology, but the strategies used by humans who work with this technology have not adequately been modified and adjusted in response to these transformations. For example, the chances that an AFIS search will produce prints with incidental similarities—i.e. that *highly similar, look-alike, prints from different sources will result from an AFIS search*—has not been sufficiently investigated or explored. This risk, as well as others, may mean that the use of AFIS introduces new concerns into the process of latent fingerprint identification, some of which may even increase the chances of making erroneous identifications. Only by appropriate and explicit adaptation to the new potential and the new challenges posed by the new technology will AFIS and other cognitive technologies produce efficient and effective partnerships.

Keywords: cognitive technology; bias; AFIS; database searches; fingerprint identification; forensic science; evidence experts; judgment and decision making.

With advances in computer technology, increasing numbers of tasks and domains that were once reserved purely for human expertise are now within the reach of technology. These technologies carry out cognitive operations and computations similar to human cognitive information processing and may thus be characterized as *Cognitive Technologies* (Dror, 2007a,b). These technologies may

† Email: i.dror@ucl.ac.uk. More information is available at www.CognitiveConsultantsInternational.com
‡ Email: mnookin@law.ucla.edu

distribute cognition in a number of ways: They can work to help and support the human expert; they can work alongside the human expert in a collaborative partnership; or the technology can play a more critical and significant role than the human.

When a cognitive technology helps and supports the human expert, cognitive operations and tasks are 'offloaded' by the expert onto the technologies (Dror & Harnad, 2008b). In this case, the technology operates in the shadow of the expert. The expert offloads some of the tasks he or she formerly carried out to the technology, which executes those tasks and provides information back to the expert. In this mode of human–technology interaction, the human expert remains clearly (and exclusively) in charge. The technology is subservient to the expert, increasing the expert's efficiency and decreasing the expert's cognitive load. For example, when a human uses a computer to store information instead of memorizing it, or uses a calculator to perform mathematical calculations, the human expert is 'offloading' tasks to technology to save time and effort. In this case, the technology offers primarily a gain in efficiency, a quantitative change, rather than qualitatively transforming what is possible.

A different mode of distributed cognition occurs between technology and human experts when they work side by side as partners. Here the cognitive tasks are divided between the human expert and the technological apparatus. In this mode, some of what the technology contributes the human expert is incapable of doing without the technology (and vice versa: the human expert also makes significant contributions that cannot entirely be offloaded to the technology). Examples of such human–technology partnerships include evaluating whether a suspect in a criminal case is a likely contributor to a mixed sample of DNA, making a clinical diagnosis partly on the basis of an X-ray, and the subject of this paper, the determination of whether a latent fingerprint found at a crime scene does or does not 'match' a source print contained in an Automated Fingerprint Identification System (AFIS), a computerized database of fingerprints.[1]

When 'expertise' is distributed between a human and a technology, it will invariably be useful to have a clear understanding of the different tasks involved within the particular domain and the relative strengths and weaknesses of the technology and of the human expert in performing their respective tasks (Busey & Dror, in press). Then by matching the specific tasks to the relative strengths of the technology and of the experts, one can hope to optimize performance in a 'divide and conquer' fashion. Over time, the distribution of cognition may change as new conceptualizations or technological shifts make new kinds of 'offloading' possible, or when experience reveals that human expertise trumps technology in ways not previously understood.

A third mode of distributed cognition involves an even greater and more meaningful role for technology and automation, while still preserving some limited province for the human. Here, the bulk of the 'important' work is done by the technological process, but the human expert still retains a degree of involvement and perhaps some limited control. In this mode, the expert operates in the shadow of the technology. For example, breath test detection for alcohol probably falls into this category; the evidentiary instruments used to evaluate the quantity of alcohol in the breath have become automated to the extent that they basically produce a result without significant human intervention. While

[1] AFIS is not one thing or one system. Rather, there are a variety of databases and numerous software products in use to enable computer-assisted searches for fingerprint identifications. Different jurisdictions have their own AFIS systems; the Federal Bureau of Investigation (FBI) uses Integrated Automated Fingerprint Identification System (IAFIS). There are multiple vendors who provide software, and vendors' search techniques and algorithms vary (though their details are generally kept confidential as trade secrets by their makers). In this paper, we employ the term 'AFIS' generically to refer to computerized pattern recognition systems for fingerprint identification.

trained human beings still play a residual role (e.g., calibrating the machines, entering and confirming basic identifying information about the test-taker, making certain the instrument is functioning within appropriate parameters, etc.), the technological instrument is making the meaningful judgements. The 'work' needed to produce a test result owes far more to the instrument's operations than it does to the relatively ministerial tasks performed by the human operator (Mnookin, 2008a).

All three of these modes of distributed cognition exist, of course, in between two extreme end points. At one extreme lies complete technological automation in which the technology replaces the human expert altogether. At this extreme, in contrast to all three modes of interaction between the human expert and the technology described above, the cognitive operations are no longer distributed at all but are solely within the control and domain of the technology. At the other extreme, human cognitive activity operates without any aid from technological tools. Here again, there is no distribution between the human and the technology; the activity is entirely conducted by the human. The three modes described above are middle grounds between these two dramatic (and often idealized) extremes.[2]

Although we have identified three points on the line between these two extremes, this entire line is more appropriately considered as a continuum. In addition, establishing the nature of distributed cognition along the continuum is complex. For example, when the computational speed of a computer so dramatically exceeds that achievable by a human, the 'mere' efficiency gains of the first mode of cognitive distribution can transform from simply a quantitative increase in speed or capacity into something qualitatively different, enabling new paradigms and new questions to be asked and answered than were previously possible. Similarly, in many cases it may be far from obvious whether the technology and the human play equivalent roles in the partnership or one is subservient to the other. Take, for example, an automated process in which the human acts as a 'fail safe' who takes action only under limited conditions when the automatic process does not suffice (e.g., the use of autopilot when an aircraft is in cruising mode). In such a system, is the human best thought of as the ultimate controller of the system, the one with the final judgement in the hardest circumstances; or should we conceive of the human as playing a rather minimal and secondary role, as in most instances his or her expertise will not even be called on? Is this a version of our second mode, an example of our third, or somewhere in between?

In many situations, the particular pragmatic context in which the human–technological interaction takes place may affect those interactions and how cognition is distributed. Consider a human who uses a global position system (GPS) system to navigate a route between two locations. If the person is in an unfamiliar city, we may think that this falls into our third mode, in which the human is 'subservient' to the GPS, following its route uncritically without exercising independent judgement. In contrast, a human using the same technology in his or her hometown may treat the GPS' suggested route as nothing more than a loose guide to be overruled at will on the basis of the human's independent knowledge and expertise. In short, these three modes—technology as cognitive servant, technology as cognitive partner and technology as cognitive driver—are ideal types in the Weberian sense, rather than literal descriptions of the range or particular possibilities (Weber, 1949).

[2] In offering this framework, we are avoiding the obvious complexities involved in trying to define either 'technology' or 'cognition'. We note, however, that depending on how expansive one wishes to be in defining 'technology', it may actually be hard to envision human cognition unmediated by technology. And of course, even with total automation, the system at issue has been designed by—and is quite possibly maintained and evaluated by—humans. It is beyond the focus of this paper to demarcate precise boundaries between the technological and the human, or even the boundaries between our three categories. Rather, we aim to frame these different modes as important points on a continuum.

The central point is that we will gain useful, and indeed necessary, insight (and perhaps avoid significant errors or even disasters) from a better understanding of the specific dynamics of human–technology interactions and the particular modes of distributing cognition within any given system. Using any technology effectively requires us to have a reasonable degree of understanding of the role of the technology, the role of the human and their interaction and mutual influence within the 'technological system' as a whole (Hughes, 1983). It is not that any point on the continuum is necessarily superior to any other, but rather that by understanding where on the continuum we are, why we are there, and the relative capacities, incapacities and relationships between the technological elements of the system and the human ones, we will greatly improve our ability to put both the technological and the human aspects to optimal use.[3]

In this paper, we specifically consider and focus on distributed cognition as it relates to a widely used technology in the domain of forensic science: AFIS. We will examine its use, in relation to latent fingerprint identification in the criminal justice system. How, if at all, has the increasing use of this technology transformed what human fingerprint examiners consider when they compare prints and reach a conclusion,[4] and how does the use of AFIS transform what they *should* consider? How do their cognitive tasks change when they 'partner' with AFIS, as compared to what was needed to make latent fingerprint identifications before computer systems and the use of enormous databases they entail? Although we focus in this paper exclusively on AFIS and fingerprinting in the domain of forensic science, the more general point regarding the need for careful consideration of both the details and the consequences of distributed cognition between technology and expert is applicable to the use of technology in other expert domains, such as medical, aviation, financial and e-learning.

Our central aim is to suggest in strong terms that these questions about the appropriate effects of AFIS on the process of latent fingerprint identification warrant significant further inquiry and consideration. We believe that AFIS *does* change in important ways the cognitive tasks in which latent fingerprint experts are engaged. Our key point is that there has not yet been sufficient consideration of either the meaning or the consequences of the new distributed cognition that AFIS offers, either the new potentialities or the new risks for error. Such understanding will enable better deployment and utilization of this technology, e.g., Dror (2006a, 2007b).

[3] Using cognitive technologies in different ways may not only create new capacities but also change both the quality and the nature of the work we do, and even more significantly, it may change us—humans—in important ways (Dascal & Dror, 2005; Dror, 2007a). The implications of using cognitive technology range from affecting how we organize information to how we think and communicate with one another. Moreover, cognitive technologies are far from an unmitigated good; we must understand their limitations, pitfalls and vulnerabilities, which may lead to a variety of negative or even disastrous consequences (Dror, 2007b). For example, will regular reliance on GPS systems inevitably reduce people's ability to use and navigate with conventional maps? Will the availability and use of automated spelling checkers which detect and fix spelling mistakes reduce human cognitive capacity for spelling? And if so, are such outcomes of cognitive technology all bad or are they advancements in human development? In this light, cognitive technology may be viewed as a new evolutionary force that works in relatively super fast time scales. Finally, the proper development and deployment of cognitive technology is critical if it is going to be effectively put to its intended use in a variety of domains. For example, with most patient monitors in intensive care units, the monitors aimed at reducing the cognitive load of the medical staff are set at such a low threshold that they 'go off' (causing false alarms) so frequently that they are largely ignored by the medical staff. These questions and examples illustrate the need to develop cognitive technology, across domains, that properly works with humans by considering the collaborations with humans (both direct and indirect, immediate and long term), and the implications of how cognitive technology distributes cognition.

[4] Throughout the paper, when we refer to the end of the cognitive process in latent fingerprint identification—the conclusion reached, whether an identification, an exclusion or a claim of insufficiency—we also include all perceptual and cognitive processes involved in examining and analysing the prints prior to reaching the conclusion.

Put simply, *the use of AFIS ought to change the way fingerprint experts conduct comparisons, and what they require in order to declare a 'match', because making identifications is simply not the same cognitive task as it was prior to the use of massive, automated computerized databases.* More specifically, we believe that insufficient attention has been paid to at least three dimensions of the AFIS–human partnership: (1) whether AFIS ought to change fingerprint examiners' decision thresholds for reaching a conclusion, (2) whether the design of the AFIS–human interaction creates new possibilities for biasing human examiners and (3) how AFIS' methods for evaluating prints interact with human methods, and what effects these combined methods might have on the probative value of a conclusion. Our first and foremost purpose in this paper is to draw attention to these questions. The effects of this new technology on the process of latent fingerprint identification deserve attention, both from AFIS vendors and from the latent fingerprint community and professional bodies. Fingerprint examiners themselves may have a great deal of practical wisdom about AFIS, its power and its limits, but this has not been formalized or used systematically. Attention is equally required from evidence scholars, statisticians and cognitive psychologists, who have largely not considered these arguments with respect to AFIS in particular or latent fingerprint examination in general.

AFIS, in a nutshell, is a cognitive technology system that compares the similarity across fingerprints. The FBI's IAFIS, one of the largest fingerprint databases, contains the prints of more than 60 million individuals (and over 600 million separate prints). These systems are extremely powerful and impressive as they can compare millions of prints in a very short time (although their performance capabilities are, like virtually all forensic sciences, rather dramatically exaggerated by CSI's glossy depiction of fingerprint matching). Over time, their capacities have grown, their speed has increased and they have been widely recognized as an extremely helpful crime-fighting technology (Klug *et al.*, 1992; Peterson & Moore, 1996).

At present, AFIS systems are used for multiple purposes, and the human–technology interaction varies depending on the purpose. For example, AFISs are frequently used to compare a particular individual's 'tenprints' (inked or scanned prints of a person's 10 fingers, taken in a controlled manner) to the tenprints stored in the database in order to confirm or discover a suspect's true identity or determine whether the individual has a criminal record (Peterson & Moore, 1996). In this circumstance, the technology has access to a great deal of information—the subject's prints are generally of high quality, having been taken under circumstances in which they can be assessed and retaken if they are not adequately clear. The system typically uses more than one of the subject's fingerprints for matching. In this environment, AFIS is thought to perform exceptionally well, as well as—and perhaps even better than—human experts.

Although there are certainly variations in procedures from jurisdiction to jurisdiction, a number of organizations have made tenprint fingerprint matching with AFIS into a fully automated process, in which AFIS itself determines whether there is a 'match' in the database (Komarinski, 2005). As long as the result reaches a predefined threshold for a 'match', the system reports the match back to the inquiring entity without any human intervention. This fully automated process mode is known in fingerprint circles as 'lights out' because with no humans involved, the lights can be turned off in the office and the system will still be able to function.

This 'lights-out' use of AFIS is, of course, an example of a technology achieving so much sophistication at what was previously a 'human task' that it replaces the human expert altogether. Some states, however, have not fully embraced a fully 'lights-out' approach even to tenprint matching, retaining a role for a human as verifier in the process. Nevertheless, it is fair to say that at least for

tenprint searches, the general trend is toward an increasing use of AFIS technology and a decreasing role for human examiners. To a greater or lesser extent, then, in the world of tenprint identifications, the cognitive processes previously assigned to humans are increasingly being offloaded to the cognitive technology.

However, a very different picture emerges in the world of latent fingerprint identification. Latent fingerprints, often collected at crime scenes, are frequently far from ideal. They frequently contain only a small portion of a print and that small portion may also be distorted.[5] The prints may contain artifacts, visible indicators on the latent print that are not associated with the pattern on the finger itself, 'noise' rather than 'signal'. Tenprints are fingerprints that are taken intentionally to generate a fingerprint image, while latents are chance artifacts—serendipitous smudges inadvertently left by individuals as they go through life. Compared to tenprints, latent fingerprints tend, as visual objects, to be incomplete, messier, less pristine and thus harder to interpret.

These circumstances make the task of comparing a latent to its source significantly more complex, requiring flexibility and the ability to supplement the minutiae detail with other types of information (such as ridge length and curvature, as well as level 3 detail, such as ridge width and sweat pores). At present, for latent prints, AFIS' capabilities are thought not to surpass, or even to meet, those of human experts,[6] and therefore no one currently advocates taking a fully 'lights out' approach to latent fingerprinting.[7]

Latent print searches in AFIS systems primarily operate in one of two ways: either latents associated with a crime scene are examined against the tenprint database for possible match candidates or the tenprints of a possible subject are examined against a database of latents from unsolved crimes. A third mode is available as well, in which a crime scene latent print is examined in relation to the database of unsolved latent prints on file. But in none of these approaches does AFIS itself determine whether or not a match exists. Instead, AFIS only provides the examiner with a ranked list of candidates that the algorithms used by the system deem most likely to match, along with a numerical score associated with each candidate on the list, which captures the AFIS algorithm's assessment of the similarity of the prints.[8] A human expert then uses the traditional methods of comparison to determine whether or not a match actually exists between the latent print and any of the possible source candidates. In these cases, the mode of using AFIS is collaborative; it is a form of distributed cognition. AFIS provides a set of possibilities to the human examiner, but it is the human examiner who determines which, if any, is an 'identification', i.e. which, if any, derives from the same source as the latent print.

[5] The FBI estimated in a study it conducted that the typical latent print included an average of 21.7% of the area of a rolled print (*United States* v. *Mitchell*, 2004).

[6] There appears, however, to have been remarkably little testing of AFIS' performance compared to examiners' performance, or even of the accuracy of examiners' AFIS-informed determinations (Haber & Haber, 2004).

[7] The National Institute of Standards and Technology recently completed a pilot study to examine the accuracy of various systems for encoding print features and searching for latent print matches automatically. The results 'demonstrated a level of performance beyond pre-test expectations' (Indovino *et al.*, 2009). More generally, the same tendency towards increasing use of technology present in AFIS tenprint processes is visible with respect to latents too, though the pace is significantly slower, and whether truly lights out systems will ever be viable is certainly both an open question (Meagher, 2009; Komarinski, 2009) and a source of potential anxiety for latent print examiners (Mnookin, 2008b).

[8] These scores are not necessarily intended by manufacturers to be taken as absolute indicators; indeed according to the documentation accompanying some AFIS systems, the relative scores across different prints may be more meaningful than the absolute scores (Cole *et al.*, 2008).

When comparisons get more and more challenging along certain dimensions, AFIS becomes ever less capable and the need for the human fingerprint expert becomes still more acute. For example, as a latent fingerprint from a crime scene contains more and more 'noise' (in the extreme case, when a number of latent fingerprints are superimposed on top of one another), then the AFIS system's automated method for minutiae detection are likely to be highly inadequate; instead, the human expert must determine and tell the computer what signals (what minutiae fingerprint features) ought to be taken into consideration. In fact, even for 'ordinary' latent prints, rather than relying on automated minutiae extraction processes, examiners typically have to provide minutiae information to the AFIS system and they even develop experiential knowledge about how best to designate minutiae within different AFIS systems or how to re-encode minutiae in order to generate a different list of possibilities (Office of the Inspector General (OIG), 2006; NAS Report 2009). Sometimes the correct match, although in the database, may not even appear on the candidate list at all; fingerprint examiners are well aware of this limitation. AFIS is also less likely to succeed when comparing a latent print from an adult to that of a young child because of the differences in pattern size that develops over the years. Such a comparison, however, would not preclude a human expert from making a comparision. Furthermore, human experts are sometimes capable of making identifications with latent prints that are deemed unsuitable for AFIS because the prints do not have the minimum quantity and quality of information required to run an AFIS search.[9]

Thus, we can see that the AFIS–expert relationship ranges greatly across the continuum we discussed at the outset of this paper. In certain circumstances, like routine tenprint searches, some jurisdictions have moved to complete automation, more or less substituting the human expert with technology (Komarinski, 2005). In other circumstances, like comparisons involving highly distorted latent prints, AFIS searches may not be of any assistance to the human. But for 'typical' latent print searches within the database, the relationship between the technology and the human takes the form of a significant collaborative partnership, with cognition necessarily and meaningfully distributed between them.

What AFIS has done in tenprint searches is streamline and made more efficient a type of search that was generally feasible even before the existence of computer searching and computer pattern recognition systems (Komarinski, 2005). In fact, efforts to classify fingerprints in order to permit searches for identification purposes date back to the late 19th century. Schemes like the Henry and Vucetich classification system produced workable methods for establishing the identity of a suspect. These methods were, overall, quite successful (Henry, 1900; Polson, 1950, 1951; Beavan, 2001; Cole, 2002, 2004). But these methods classified individuals' prints using all ten fingerprints (e.g., by assigning whorls different values for specific finger positions followed by accounting for loops and arches in the remaining finger positions), and therefore were of little assistance in identifying the source of a latent print found at a crime scene. Efforts for single print classification systems did emerge but were never as effective as tenprint classification schemes, especially in relation to a large number of prints (Cole, 2002, 2004; Battley, 1931).

In other words, with tenprint searches, AFIS has made the process of identification dramatically faster and more efficient than it used to be, but it did not enable an altogether new kind of search

[9] The expert, in such cases, may still find creative ways to bypass the AFIS restrictions and use a database search process nevertheless (such as creating 'phantom minutia').

process. By contrast, with latents, prior to AFIS, searching among large numbers of individual prints in hopes of making a 'cold hit'—making an identification among a large population of possibilities on the basis of a fingerprint match alone, without other evidentiary information to narrow or delimit the range of 'possible' sources of the print—was monumentally laborious and certainly was never possible at anywhere close to the scale achievable today (Cole & Lynch, 2006; Komarinski, 2005). It is AFIS' capabilities that have made it possible to search relatively easily for 'cold hits' in fingerprint cases.

Thus, in tenprint searches cognitive processes are 'offloaded' from the human onto the technology; in latent print searches, by contrast, the technology has created genuinely new investigatory possibilities, enabling inquiries that would have literally been impossible prior to the human–technology partnership. The combination and integration of AFIS and human expert knowledge is what has made cold hits on latent prints substantially more feasible. While there is no aggregate data on how many cold hits there have been as a result of AFIS, it is clear that there are many such cases. The FBI reported in 2006 that roughly 1,200 latent searches had resulted in identifications since IAFIS went online in 1999—a small proportion of the total searches conducted, but vastly more 'cold hits' than would have been possible or even imaginable without computerized searching (OIG, 2006).

A new technology may not only change what humans can achieve but may also change the methods or the meanings of the actions in which they were already engaged. If humans do not adapt appropriately to the new technology, understanding how it transforms both the possibilities and their practices, the human–technology collaboration will fall short of its potential. When a new technology and human experts work together, collaborating and distributing the cognition involved in fingerprint identification, *what are the challenges and risks of using the technology? Specifically, how, if at all, does the use of AFIS affect how a fingerprint examiner should analyse whether or not two prints come from a common source? How do, or ought, fingerprint examiners' tasks to change in light of the new technology? How should one assess the probative value of the evidence produced through this transformation of the human–technology collaboration vis-à-vis latent fingerprint identification?* These are the questions we wish to begin to explore in the remaining portion of this paper.

There is no doubt that the introduction of AFIS in latent fingerprint identification has been in many ways a significant success. It is the consequences of this quite impressive partnership that we wish to probe in the remainder of this paper. Its achievements notwithstanding, we believe that by failing to think through all the consequences of the use of AFIS for latent identification, not only is latent fingerprint identification not living up to its full potential but also that *the chances for incorrect identifications have increased*.

More specifically, we are troubled that in the wake of AFIS, some essential elements of how experts evaluate fingerprints and their decision-making process that ought to have received sustained attention and modification have nevertheless remained unchanged. We wish to describe three ways in which, in our view, latent fingerprint examiners and the forensic science community have not yet fully thought through the consequences of their use of AFIS. First, we believe that the use of AFIS ought to change fingerprint examiners' decision threshold for when to declare a match. Second, the use of AFIS creates new possibilities for biasing the human examiner, and these issues of possible context bias need more attention and research. Third, the question of how AFIS's methods relate to human examiners' methods, and what effect these combined methods have on the probative value of a conclusion, also deserves more sustained attention than it has yet received.

1. Standards for identification

Before the use of AFIS, experts would generally compare a latent print with a quite limited number of tenprints. Most of the time they would compare the latent to the prints of individuals already suspected of a possible connection to the crime, typically a handful of suspects. Even if they were comparing the latent to a larger possible pool, say to a few hundred or even a few thousand tenprints, these are still relatively small numbers, orders of magnitude smaller than the vast quantity of tenprints available for searching in our current AFIS databases. Given these relatively small numbers and the high degree of variability of prints across individuals, the likelihood that by pure chance one of the comparison prints would be extremely similar to the latent print from the crime scene was very low. AFIS changes this story dramatically. Latent prints can now be compared to many millions of prints stored in huge databases. In such cases, the chances of finding by pure coincidence a look-alike print, a print originating from another person but that is nevertheless extremely similar to the latent print, is much higher than when comparing the latent print to just a few dozens, hundreds or even thousands of prints prior to the introduction of AFIS (Dror *et al.*, 2005; Cole, 2005; Mnookin, 2004).

This point ought to be, in a sense, an obvious one. When database size increases, the chances that some print in the database will bear a high degree of resemblance to the latent in question also goes up. This is not a new problem. It is analogous to the issue of 'adventitious', or chance, matches in DNA: suppose the odds that a person selected out of the population at random will match a given DNA sample found at a crime scene is determined to be 1 in a million. If the database is sufficiently large, it is likely that someone in the database will be a 'random' match—that is to say, someone who truly *does* have the same DNA markers at the tested loci, but is nonetheless *not* actually the person who left the biological sample at the crime scene.[10] Similarly, as AFIS databases get larger, the chances of finding a highly similar print from a different source must necessarily increase.[11]

[10] This occurred, e.g., in the much publicized case of Raymond Easton. Despite his physical incapacity and solid alibi, he was accused of a crime on the basis of a DNA match when tested at six loci. Although the random match probability was calculated at 1 in 37 million, Raymond Easton's DNA did not in fact match the crime scene sample when additional loci were tested (Mnookin, 2001).

[11] It is important to note that questions surrounding interpretive issues in 'cold hit' DNA databases have received a great deal of attention (Kaye, 2009). Some commentators have argued that when DNA is found through a 'cold hit' database search, the probability that is reported to the factfinder that a random person would match the crime scene DNA should be adjusted to correct for the size of the database and the possibility of a coincidental match (National Research Council, 1996). Many other statistical experts, however, have strongly disagreed, arguing, in essence, that a unique DNA match resulting from a 'cold hit' search not only provides the information that someone *did* match the source DNA but also *excludes* as possible DNA matches everyone else who is in the database. According to this alternative and conflicting point of view, the information gain resulting from being able definitively to exclude everyone in the database (except the one match) as a potential source of the DNA more than offsets the need for an adjustment in random match probability because of the database size. Therefore, according to this view, correcting for database size when reporting the DNA random match probability resulting from a database search is statistically inappropriate and/or legally unnecessary (Komarinski, 2009; Balding & Donnelly, 1996; Donnelly & Friedman, 1999; Balding, 2002). *Regardless of ones position about this issue vis-à-vis DNA databases, the argument clearly does not apply in the AFIS context.* There is a critical difference between AFIS databases and DNA databases: barring testing error, one can be confident that a DNA database search will turn up any profile in the database that in fact matches the source DNA at the tested loci. If one and only one match is reported, this establishes that *no other profiles in the database match at the loci in question*. Everyone else in the database can therefore be positively excluded as a potential source. By contrast, with AFIS searches, even if a print from the same source is in fact in the database, it is not certain that it will be suggested by the AFIS algorithms as a potential source print. In one experiment,

Well before DNA evidence existed, the California Supreme Court recognized this same possibility of the danger of 'look-alike' matches, in its appendix to *People* v. *Collins*, a famous case regarding statistical evidence that still makes an appearance in many evidence casebooks. In the opinion, the court outlined mathematically the fact that with a large enough population pool of possible suspects, it became quite likely that some other couple, in addition to the perpetrators, would also have all the various personal characteristics linked to the perpetrators (*People* v. *Collins*, 1968). We can see the point in a more casual form as well. If a house is burglarized by a one-armed man with a tattoo, and later that same evening, a one-armed man with a tattoo is found merely blocks away, that may be considered fairly strong evidence suggesting he is the perpetrator, particularly if having one arm and a tattoo are quite rare in that neighbourhood. By contrast, if a one-armed man with a tattoo is noticed the next day 2,000 miles away, our confidence that the person with those characteristics was linked to the crime would appropriately be significantly reduced. In the immediate vicinity of the crime, it is likely that there are very few one-armed men with tattoos. Across the entire country, there are likely to be a good many more. The breadth of the search parameters necessarily affect the meaning and probative value of the evidence that is found.

What AFIS does, in essence, is analogous to the shift from looking for the one-armed man with a tattoo only in the immediate vicinity of the crime to searching for him on a national scale. But despite this transformation, fingerprint experts have not changed and adapted their criteria for determining what counts as an identification as a result of AFIS. Their task and their methods—determining whether there is sufficient evidence of similarity between the latent print and the exemplar used for comparison—have remained fundamentally the same, whether the exemplar to which they declare a match has come from an individual who is suspected of the crime for other reasons or the exemplar results from computerized pattern recognition software that has determined that these prints are among the most similar to the latent source print of all the millions of prints in the database.

And yet surely comparing very similar prints to one another is a riskier exercise than comparing quite different ones. When experts compare prints, the likelihood of making a wrong match, an erroneous identification, must be considerably higher when comparing two very similar prints than when comparing two very different prints (Haber & Haber, 2004). This is quite obvious and supported by scientific research (e.g., Ashworth & Dror, 2000). It is an artifact both of the relative similarity of the patterns being compared and of the human cognitive architecture and processes involved in pattern matching (see Ashworth & Dror, 2000). Therefore, with the introduction of AFIS and the increased likelihood it entails of finding very similar patterns by pure chance, human experts should modify and adapt their decision-making threshold when comparing prints that are suggested by AFIS. They should require *more* evidence of similarity when making an AFIS match than they would require elsewhere.

A fingerprint examiner might think that because he or she declares an identification only when he or she can confidently 'individualize'—attribute a latent print to a particular individual, to the exclusion of all others in the world–he or she need not be concerned with database size.[12] Our point

e.g., an AFIS system failed to report the matching print in the database as one of the 10 potential matches 18% of the time (Cole *et al.*, 2008). To put the point differently, AFIS' algorithms are both more complex and less accurate vis-à-vis latent prints than DNA database searches. Therefore, the informational gain achieved in the DNA context from the ability to exclude all the non-matches with confidence as potential sources of the DNA has no parallel in the AFIS context at present.

[12] Whether fingerprint examiners can legitimately claim to individualize (e.g., see such a claim asserted in Barnes, 2009, p. 17) has been a source of significant contention in recent years, see, e.g., NAS, 2009; but this issue is not relevant to our paper.

is that the quantum of information that latent fingerprint examiners believe necessary to individualize was determined in a pre-AFIS era, and now, with introduction of AFIS technology this quantum may need to be revisited. Whatever degree of 'sufficient similarity' (Dror, in press) was necessary to be confident about an identification before AFIS made it possible to search millions of prints to find the most similar among them can no longer be presumed to suffice. It is in this sense that decision-making thresholds and criteria ought to increase because of database size. However, this has not happened, at least not in any formal way; the same decision-making criteria and threshold are prescribed regardless of whether AFIS is used or not. This is—or at least ought to be—a significant concern.[13]

To be more precise, whether or not examiners actually do continue to use the same decision criteria is difficult to assert with confidence because declaring an identification by a fingerprint expert is a subjective judgement. It is certainly conceivable that in actual practice (consciously or not), some fingerprint examiners do require more evidence of similarity when assessing prints generated by AFIS than when assessing prints to which they are directed in other ways.[14] But there are no formal guidelines whatsoever on this issue or any official or professional injunctions to take this concern onboard. Nothing, e.g., in the Latent Process Flowchart (NIST) or in the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) guidelines suggests that decision thresholds should be higher for database searches.[15] Indeed, even in the National Academy of Sciences' generally thorough discussion of a vast array of concerns surrounding forensic science, there is no discussion about the risks large databases may pose for generating incidental similarities nor any recognition of the need to modify decision thresholds as the scale of possible matches being searched increases. There is, in fact, an entire chapter on AFIS in the report, but nearly all the discussion in the chapter focuses on questions of interoperability across different AFIS's, rather than considering how AFIS might appropriately (or inappropriately) affect match thresholds (NAS, 2009). Nor are there any published discussions by practitioners or AFIS vendors that systematically explore this issue.

In fact, such adaptations of criteria are standard practice in science. Take, as one example, the widely understood need to adjust the criteria for significance levels in multiple testing via a Bonferroni correction (Bonferroni, 1936). When examining statistical significance in scientific investigations, a P value of 0.05 is used as a threshold; if the P value is over 0.05 the results are deemed insignificant, whereas if it is less than 0.05 the results are statistically significant. However, this alpha level threshold can get inflated as a result of multiple testing because each test has an incidental chance (as small as it may be) of finding statistical significance by pure coincidence.

[13] This issue is made more acute by the practice of fingerprint examiners to describe the evidence in only three ways: as an individualization—a claim that this latent came from a given source and cannot possibly have come from anyone else; an exclusion—a claim that this latent cannot possibly have come from the source in question—or as insufficient for reaching any conclusion. There is, in other words, no possibility for declaring a 'probable' match or asserting an identification but with a reduced level of confidence.

[14] Perhaps print examiners might increase standards out of an intuitive sense of the heightened dangers of false positives with respect to database searches. Or, alternatively, they might increase the standards as an act of mistrust in the technology, which could lead to treat an AFIS' suggestions with extra suspicion and care. Of course, it is also conceivable, at least as a purely theoretical matter, that trust in the technology could lead an examiner to operate with a lower decision threshold as well (similar issues arise in a human-only process whereby one human verifies the conclusion of another, and the verifiers knowledge—or belief—in the first examiner's ability can affect the level of scrutiny and check carried out during verification).

[15] SWGFAST is a professional body that provides guidelines and standards for fingerprint examination; see www.swgfast.org.

As statistical testing is conducted more and more times, the chances of encountering a result that appears significant but is actually a coincidence increases. Thus, Bonferroni (or other) corrections are necessary countermeasures to correct for this risk (Benjamini & Hochberg, 1995; Shaffer, 1995). This adaptation for multiple testing, which is standard scientific practice, is one example of the need to adjust criteria depending on the number of comparisons and tests being conducted. In fingerprint identification, with the huge increase in the 'scale' of testing introduced by AFIS, the human expert's evaluation of the AFIS output must be changed to reflect the higher chance of coincidental similarities. The problem is that despite the widespread use of AFIS, there has not yet been scientific consideration or the development of guidelines as to whether (and how) human experts should adapt their decision threshold.

Remember that the very aim of AFIS, its purpose, is to seek out the most similar prints in the database by testing and comparing each stored print against the latent exemplar from the crime scene. The whole point is to find whatever prints most resemble the latent according to the algorithms and search parameters provided, whether or not the actual source of the latent is even in the database at all. AFIS must, *by design*, increase the chances that the examiner will be presented with quite similar look-alike prints, as compared to those prints presented if suspects were identified through traditional investigative techniques rather than AFIS. Perhaps, AFIS is therefore not only an acronym for Automated Fingerprint Identification System but also should equally stand for *Automation Fuels Incidental Similarities.* Of course, it is not simply that *automation* fuels the possibility of incidental similarities. More precisely, it is the transformation in scale, the vast increase in the quantity of information being searched and tested. But that vast increase in scale is itself possible only because of the advances in cognitive technology: human examiners acting alone could not possibly search through millions of fingerprints.

These concerns are not simply abstractions. The erroneous identification of Brandon Mayfield as the Madrid bomber was due, at least in part, to the coincidental similarity between one portion of one of his fingers and the portion of the latent print submitted to the databases (e.g., Stacey, 2004). In this case, four separate examiners—including one hired for the defense—concluded in independent examinations that Mayfield's print matched the print found at the crime scene. All four turned out to be wrong. The print was subsequently attributed to an Algerian residing in Spain, Ouhnane Daoud.

What was the role of AFIS in this erroneous identification? First of all, the highly similar print was found precisely through a huge database search. The larger the database, the greater the chances of finding a highly similar print that belongs to someone other than the source of the latent. Indeed, the Office of the Inspector General, which investigated and wrote a detailed report on the Mayfield case, even recognizes at one point that part of the problem is precisely the scale of AFIS searches and the possibility that they will generate incidental similarities:

> The Mayfield case demonstrates the potentially misleading power of IAFIS. . . . IAFIS is designed to find not only the source of the print (if it is in the database), but also the closest possible non-matches. In other words, although no two people have identical fingerprints, there are some that may be sufficiently close to confuse an examiner dealing with a latent of imperfect clarity. An IAFIS search of a huge database is designed to find those prints most likely to confuse an examiner. The likelihood of encountering a misleadingly close non-match through an IAFIS search is therefore far greater than in a comparison of a latent print with the known prints of a suspect whose connection to a case was developed through an investigation (OIG, 2006, p. 137).

But elsewhere in report, the OIG shies away from its own conclusions. They explicitly reject as an explanation what they call 'excessive faith in IAFIS technology' (OIG, 2006, p. 189). While they are correct that the examiners did not simply accept the IAFIS results without making their own independent comparisons, a key problem was precisely that they failed to understand how the scope of the database search needed to affect their decision threshold.

Moreover, the report determines that a key cause—not the only cause to be sure, but a central one—was the 'unusual similarity' between the two prints (OIG, 2006). The report emphasizes how extraordinarily rare it believes it to be that two prints share so many similarities:

> Even taking into account the ambiguity as to whether particular features were ending ridges or bifurcations, the correspondence of 10 Level 2 details in prints from different sources, in sequence and with consistent ridge counts, is an extremely unusual event. Although the OIG found no exhaustive or systematic study of the rarity of such an event, anecdotal reports in the literature of similar fingerprints from different sources suggest that nobody has yet demonstrated more than eight or nine Level 2 details in sequence from different sources, even using prints that were artificially cropped to omit dissimilarities (footnote omitted) (OIG, 2006, p. 136).

This, however, is reaching a premature conclusion about a matter that should instead be an important empirical question. Prior to the use of computerized databases, there is no doubt that the experience of latent print examiners suggests that two prints from two sources as similar as Daouad's and Mayfield's was a tremendously unusual event. But prior to AFIS, no one was searching many millions of prints precisely for similarities. There is virtually no research or carefully documented evidence to answer the question of how often AFIS might offer up a suggested print that is sufficiently similar to the latent print in question as to mislead experienced examiners into believing it is a match. How often might AFIS candidates 'fool' an expert in this way? This simply has not yet been adequately asked or researched, let alone countermeasures employed to protect against the risk. But the Mayfield case certainly illustrates that concerns about AFIS and the appropriate decision threshold for declaring an identification are not merely academic (Dror *et al.*, 2005; Mnookin, 2004).

2. AFIS and the possibility of bias

A second set of concerns about AFIS relate to the ways that the use of AFIS might introduce new biases into the work of latent fingerprint examiners. As we have already described, AFIS does not itself make any final determinations of identifications in complex latent print matching. Rather, AFIS works in a collaborative and distributed cognitive mode with the human expert: AFIS searches through a huge number of prints stored in a database and compares them to the latent print. Then AFIS provides the human expert with a list of most likely candidates for a match. It then becomes the human expert's job to examine *these* candidates and to determining which one (if any) is an identification.

The introduction of bias may occur because of the ranking and score AFIS gives to each print on its list of 'possibles'. The human experts may examine the potential matches not purely by their objective data, but be influenced by where they appear in the list that AFIS provided and the scores AFIS assigned to each of the candidate prints.

If the examiner considers the AFIS ranking in making his or her own determination, is this actually a bias? At least until such time as we have a much more formal approach to the assessment of latent fingerprint identification, the answer to this question must be 'yes'. The ranking creates the possibility that the human expert may not examine and weigh the data by its bottom-up objective indicators but rather, may perceive and evaluate the print partly using the top-down AFIS information. These influences could unintentionally and without awareness bias the expert's decision making, causing them to too easily believe they have found a match because AFIS has for example ranked a particular print 'first' (see Dror *et al.*, 2005; Dror & Rosenthal, 2008; Haber & Haber, 2004; Risinger *et al.*, 2002).[16]

Imagine what would happen if the scores and order of an AFIS output were fabricated and modified so that the top candidate with the top score was moved to the bottom of the list and given a lower score, while several of the lower-scoring potential matches were given higher scores and moved up the list. Would providing this fabricated information affect the human expert? At one level, the answer is clearly 'yes': it would, at a minimum, change the order in which the expert conducts his or her comparisons. At present, experts usually start their assessments and comparisons with the top AFIS matches, beginning with whatever print is ranked first and has the highest score. And indeed, among those cases in which examiners find an identification through AFIS, the top-ranked candidate is in fact the most likely candidate to be matched by the human expert; according to the FBI, when a match is found using IAFIS, the source of the latent print is the top scoring IAFIS candidate 80% of the time (OIG, 2006). So at a minimum, the fabricated information would affect the examiners' order of business.

However, the more important question is whether this modification in order and AFIS scores would have an effect not on the order in which an examiner compared the latent to the possibilities, but rather, on the actual decision and conclusion reached by the human expert, i.e. if any print matched, and if so, which one. This is an important research question with empirical results that have implications for how to best design the AFIS technology and also how to determine the best practices of the human experts (Dror & Shaikh, 2005a,b). If the order of ranking and the numerical scores were found to affect outcomes, this would suggest that this information does have potential biasing effects on experts.

Of course, if AFIS rankings are in fact appropriately correlated with the likelihood of an actual identification, we might think it appropriate for the human expert to give the ranking and scores some independent weight in their analysis. But this must be done explicitly, in an appropriate way and to the right extent (see footnote 13). From a Bayesian perspective, e.g., if the AFIS ranking were understood as a separate piece of evidence from the comparison undertaken by the human expert, each could appropriately affect the likelihood that the identification was in fact correct and the evidence of the two combined might be more powerful than either taken alone. Moreover, it might well be the case that the fingerprint expert (rather than, say, a jury) was better positioned to combine those two pieces of evidence (the AFIS result, rank and score, and the independent comparison by the expert) into a conclusion about the likelihood that the latent at issue came from a particular source. But at present, fingerprint evidence is not structured in a way that would permit such calculations or such conclusions. At least for the moment, there is neither the data nor the inclination to provide meaningful probabilities that a particular print does or does not come from

[16] It is important to recognize, though, that AFIS information might justifiably be used to facilitate the human expert evaluation, if the results explicitly and intentionally affect the examiner, and only to a controlled and appropriate extent (see, e.g., Koehler, 1993).

a given source (e.g., Stoney, 1991; Cole, 2002; Zabell, 2005; Mnookin, 2001, 2008b; NAS, 2009) though there are some promising statistical methods for developing likelihood under development (e.g., Neumann *et al.*, 2007).

Another source of bias that AFIS may introduce is the possible contamination of the ACE-V methodology for fingerprint identification. The first stage in applying the methodology involves Analysis, followed up by Comparison, before Evaluation takes place (see Dror, in press, for the importance of applying this methodology in a linear fashion). However, with AFIS, the examiner retrieves a list of candidates and then screens the suggested prints, and many are rejected after only a cursory glance.[17] Thus, an examiner often quickly and easily eliminates a number of proposed AFIS suggestions as implausible. Only after finding prints that survive this quick preliminary scrutiny does the examiner begin the detailed, formal ACE-V process. And yet at this point, the examiner has, de facto, already engaged in an initial evaluation and deemed the print in question 'plausible' as a potential match. It is certainly possible that this early scrutiny could have biasing effects, in that it could possibly 'prime' the examiner to have a gut feeling that he or she is dealing with an actual identification. Thus, when examiners systematically apply the ACE-V methodology, as they begin the Analysis and Comparison stages, they have already had potentially contaminating and biasing influences deriving from their preliminary, superficial evaluation of the AFIS output. The point is that any interaction with cognitive technology may affect human cognition, and one must give careful attention and consideration of how such technology–human distributed cognition may influence perception and decision making.

3. The appropriate probative value of an AFIS match

As we just suggested, the question of how the combined evidence of AFIS and the human examiner ought to be 'weighed' is quite complex. To what extent do AFIS' methods[18] more or less replicate human methods of comparison? To what extent are they based on substantially different data or information? And how does the similarity or difference of these methods have an impact on how we understand the probative value of an identification made via an AFIS–examiner collaboration?[19]

We can ask these questions in several ways. We might for example, ask the following: if a human expert had infinite time and was asked to select those 10 or 20 prints from a database most similar to

[17] AFIS systems typically rely upon the distance between minutiae for selecting similar prints and generally do not identify or consider the overall ridge pattern of the image, e.g., whether it is a loop or a whorl. As a result, a human examiner may see that a suggested print has a different overall ridge pattern than the latent and hence instantly know that the two prints cannot have a common source, regardless of the extent to which they may contain similar minutiae.

[18] We use the term 'method' (and criteria) to designate the data, information and characteristics used by the algorithms for comparison, as well as to refer to the algorithms by which the compute program looks for candidate prints. By using the word 'method' here, we do not mean to be referring to the so called ACE-V method of fingerprint comparison.

[19] These differences have important implications across the board. For example, it is critically important to be able to classify latent prints and comparisons to various levels of difficulties (as each level may encompass, e.g., different levels of error rate associated with that specific difficulty). One can classify difficulty by the complexity of the actual prints/comparison, but the difficulty is highly dependent on the algorithms used to process the prints. What counts as difficult for an AFIS may be easy for a human expert (or visa versa), as they use different comparison processes. Thus, the similarity or differences of how AFIS and human experts go about comparing fingerprints has far reaching implication for a number of issues, such as the classification of fingerprint difficulty. In this paper, however, we focus on how these similarities and differences may impact the value of the combined collaborative work of AFIS and the human expert and on how it affects the mode of collaboration and distributed cognition between them.

a particular latent, how much overlap would there be between this humanly produced list of similar prints and those provided by a given AFIS? The more similar the two lists were likely to be, the more it would seem that whatever methods were being used by AFIS and the examiner, they were in the end, selecting for the same metrics of similarity. By contrast, the more divergent the two lists, the greater the gap between how the computer algorithms understand and operationalize 'similarity' and how human latent print examiners do.

Another way to ask this set of questions is to focus on whether AFIS and the humans are using the same 'data' in their analysis. At a gross level, the answer is obviously 'yes': they are each using the informational content of the prints themselves. But as we analyse the question in a slightly more fine-grained way, the answer almost certainly becomes 'no', or 'not completely'. AFIS might rely more on measuring the distance between minutiae and less on ridge paths, sequences and configurations than a human print examiner. A human examiner might use level 3 data—microfeatures such as sweat pores or the width of particular ridges—whereas this information is not typically used by AFIS algorithms at all. AFIS might focus more on the distance between minutiae than on the other aspects of their relationship to one another than do examiners. It is understood, for example, that:

> There is an important distinction between the IAFIS encoding process and the analysis phase of the ACE-V process . . . To encode a print for IAFIS, an examiner utilizes only part of the information that is collected during the analysis phase—specifically, the location and orientation of the selected minutiae. Among other things, the encoding process does not utilize information about the complete ridge path between points, and does not utilize Level 3 details. Nevertheless, as to the encoded points, the encoding record does reflect the examiner's contemporaneous analysis at a stage prior to the introduction of any possible bias as a result of comparison to an exemplar print (OIG, 2006, p. 119)

Thus, while it is clear that the human examiner may use different and additional information from that provided to and utilized by AFIS, how to unpack the degree and the importance of these differences is far from transparent.

To the extent that the human examiner's reliance on particular print details or a particular methodology mirrors the computer process, the expert is engaging in a form of *replication* and *redundancy* of AFIS' own methods. By contrast, to the extent that the human expert relies on information not taken into account by the computer algorithms, or to the extent that the expert's specific matching methodologies differ from the computers, then the human expert's conclusion about whether the print matches is not simply replication, but rather, offers 'new' and perhaps *cumulative* evidence in favour of or against the inference that the two prints come from a common source (see Schum, 2001, for a general discussion on redundant and cumulative evidence in court).

To put the point differently, whether or not the human expert is using distinctive information and methodologies as compared to the computer ought to have some effect on how the human expert (and eventually, if the case goes to trial, the factfinder) assigns probative value to the computer–human collaborative conclusion regarding the fingerprint. The candidate match was not randomly selected by AFIS, but rather was selected precisely because of its similarities according to certain search criteria. If the expert's analysis makes use of other information from the print that was not used by AFIS, then this information has much more value for establishing that the similarities between the prints are *not* incidental than if AFIS *did* use this same information in its analysis.

DNA once again provides a useful analogy. Imagine that a crime scene sample corresponds to a profile in the database when eight loci are tested. Now imagine a second test of the two samples: if that second test looks at those same loci a second time, it may give us greater confidence that the test was done accurately, but it does not actually provide new information or increase the overall chances (apart from whatever chance was associated with the possibility of a technical testing error not likely to be replicated in retesting) that the crime scene sample came from that person. By contrast, if the second test looks at four additional loci and finds that they also match, that provides separate and convergent—or, in Schum's terms, cumulative—evidence in favour of the hypothesis that the crime scene sample came from that source. And the same point can be made with non-scientific evidence as well. Double checking that the suspect indeed has a tattoo and one arm is one thing; finding that the one-armed tattooed man being held as a suspect shares some additional characteristics in common with the perpetrator is another, and more powerful, item of evidence.

So, are human fingerprint experts who engage in comparisons more or less replicating the work of AFIS, or are their conclusions better thought of as providing a form of converging evidence? While we are not in a position to satisfactorily answer that question, the point is precisely that it *is* an important question. It is quite clear that they are doing some of each, but to what extent, and in what ways, is, we believe, something to which more attention should be paid. To put it differently: the more that AFIS is, from the perspective of the examiner, a kind of 'black box' whose internal operations and methods are not much understood by examiners themselves, the less potential there is to aggregate the evidence provided by AFIS with the conclusions of the expert to generate even more confidence in their combined probative value. There is nothing inherently wrong with treating AFIS as a 'black box', using its suggestions without fully understanding the methods by which it was produced (Mnookin, 2008a, in press). But the more that AFIS is treated in this manner, the less information we have about whether examiners might be engaging in a kind of implicit double counting in their own analysis and the less well we understand whether the examiner's conclusion is primarily a kind of replication or might instead be a powerful piece of converging evidence suggesting identity.

Whether or not AFIS and the human expert are making similar inquiries also affects the mode of collaboration and distributed cognition. At the outset of the paper we discussed one mode of distributed cognition in which the technology is subservient to the human expert, and where the expert offloads their processes onto cognitive technology. We would be closer to that first mode if in fact AFIS uses similar comparison processes as the human. However, if AFIS's methodologies for determining similarities differs significantly from the human expert (not only in the quantity of comparisons, where it obviously does differ, but also in qualitatively meaningful ways), then the human–technology mode of interaction is not merely offloading, but rather constitutes a collaborative interaction of partnership, that of the second mode described in the first section of this paper.

More generally, the human expert's analysis of any latent print ought to be very different depending on whether the latent fingerprint examination, as a whole, is a one-stage process or a two-stage process. By one-stage we mean that the latent and a small set of comparisons are given directly to the expert on the basis of other investigative, non-fingerprint evidence, whereas a two-stage process occurs when there are two stages to the fingerprint search; first through AFIS, and then a second run-through, by examiners themselves. In the one-stage process, the prints that the examiner looks at have been selected at random, not from an overall evidentiary perspective, but

in the sense that the *fingerprint characteristics* themselves had nothing to do with their selection; they are no more likely to be similar to one another than any set of prints taken from the population at random. By contrast, in the two-stage process, the set of prints for comparison was based on AFIS's comparison of the latent fingerprint itself to those in the database. Thus, in this case the set of prints are *not* randomly chosen, in the sense that the fingerprint characteristics themselves were indeed *the* determining factor for their selection. For the AFIS two-stage process to work optimally, the analysis in stage two by the human expert should be informed and influenced by the method used by AFIS for the initial selection.[20] The fingerprint examiner should understand the extent to which he or she is replicating AFIS' selection/evaluation method and when he or she is engaging in a different selection/evaluation method that may provide 'new' evidence for a match vis-à-vis what AFIS examined. Examiners' lack of sufficient knowledge to enable this kind of coordination impedes the technology–human collaboration. In part, though not entirely, this is due to the commercial confidentiality of the specific algorithms used by various AFIS vendors.

4. Summary and conclusions

The implications of using technology with human experts are wide in scope. We have discussed how the use of large databases and multiple comparisons requires adaptations and adjustments in the threshold criteria for making an identification. We have seen how AFIS may generate new dangers of bias for examiners. We have considered how our evidentiary assessment of matches made through the human–technology collaboration AFIS offers has the potential to be even more powerful through a more nuanced understanding by the humans of the operations and processes used by the technology. In all these ways, the effects of the new technology on the human part of the equation have not yet been fully considered or adequately understood.

In this paper, we have certainly not covered all the possible issues, challenges and problems associated with cognitive technologies such as AFIS.[21] But we have, we hope, illustrated with a few concrete and practical examples that for technologies (and especially cognitive technologies) to work to their full potential one must consider carefully and understand how cognition is distributed between the technology and the human being, the nature of the particular tasks and the nature of the human–technology collaboration. Using this holistic cognitive approach not only can we reduce the chances of making errors, but we can also increase our odds of being able to use

[20] In contrast, in the one-stage method, the traditional investigative information used for initial selection may act as a biasing factor in the expert's analysis and conclusion. However, this is a different type of bias than the one we are discussing in this paper; its cognitive origins are different, it requires different countermeasures and we deal with this type of bias elsewhere (Dror, 2005, in press; Dror & Charlton, 2006; Dror *et al.*, 2006; Dror & Rosenthal, 2008).

[21] For example, collaborative partnerships with technology change the nature of the work conducted by the human element of the partnership. This often means that the humans may need to use a whole new set of skills and aptitudes in their work (just think of how the skills needed to be an architect or interior designer have changed over the years as computer-assisted drafting replaced human drafting of plans). These changes need to modify selection criteria for recruitment as well as training, in order to better match experts to the changed set of required skills and abilities. With respect to latent fingerprint experts, our sense is that selection and training criteria have changed rather little since the introduction of new technologies. Present day training includes specific and technical knowledge on how to operate the technological apparatus but does not address the fundamental cognitive issues relating to the array of new skills and aptitudes needed to work in a highly technological cooperative and cognitively distributed environment.

the technologies in as effective a way as possible, by understanding their potential, their limitations and the complex ways in which they affect what is required of the human beings that depend upon them.

Acknowledgements

Many valuable comments on earlier versions of this manuscript have been helpful to us in considering the issues and how best to present them. For helpful comments, we want to thank Eric Buel, Lauren Cooney, Norman Fenton, David Kaye, Jay Koehler, Stephen Meagher, Joe Peterson, Michael Risinger, Norah Rudin, John Vanderkolk, Kasey Wertheim, Arie Zeelenberg and an anonymous reviewer for *Law, Probability and Risk*. We also want to thank a number of additional people who provided us with useful comments but asked not to be named. Special thanks to those who provided critical comments and criticisms, some of whom will no doubt continue to disagree with views and opinions expressed in the manuscript, for which the authors bear sole responsibility.

REFERENCES

ASHWORTH, A. R. S. & DROR, I. E. (2000) Object identification as a function of discriminability and learning presentations: the effect of stimulus similarity and canonical frame alignment on aircraft identification. *Journal of Experimental Psychology: Applied*, **6**(2), 148–157.

BALDING, D. J. (2002) The DNA database search controversy. *Biometrics*, **58**, 241–244.

BALDING, D. J. & DONNELLY, P. (1996) Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Sciences*, **41**, 603–607.

BARNES, J. G. (2009) History. In A. McRoberts (Ed.) *Friction Ridge Sourcebook*. NIJ Press: Washington, DC.

BATTLEY, H. (1931) *Single Finger Prints, a New and Practical Method of Classifying and Filing Single and Fragmentary Impressions*. Yale University Press/Bureau of Social Hygiene: New Haven.

BEAVAN, C. (2001) *Fingerprints: The Origins of Crime Detection and Murder Case That Launched Forensic Science*. Hyperion: New York.

BENJAMINI, Y. & HOCHBERG, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Serie B*, **57**, 289–300.

BONFERRONI, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

BUSEY, T. & DROR, I. E. (in press). Special abilities and vulnerabilities in forensic expertise. In A. McRoberts (Ed.) *Friction Ridge Sourcebook*. NIJ Press: Washington, DC.

COLE, S. (2002) *Suspect Identities*. Harvard University Press: Cambridge, MA.

COLE, S. A. (2004) History of fingerprint pattern recognition. In N. Ratha & R. Bolle (Eds.) *Automatic Fingerprint Recognition Systems*. Springer: New York.

COLE, S. A. (2005) Less than zero: accounting for error in latent fingerprint identification. *Journal of Criminal Law & Criminology*, **95**, 985.

COLE, S. A. & LYNCH, M. (2006) The social and legal construction of suspects. *Annual Review of Law and Social Science*, **2**, 39–60.

COLE, S. A., WELLING, M., DIOSO-VILLA, R. & CARPENTER, R. (2008) Beyond the individuality of fingerprints: a measure of simulated computer latent print source attribution. *Law, Probability and Risk*, **7**, 165–189.

DASCAL, M. & DROR, I. E. (2005) The impact of cognitive technologies: towards a pragmatic approach. *Pragmatics & Cognition*, **13**(3), 451–457.

DONNELLY, P. & FRIEDMAN, R. D. (1999) DNA database searches and the legal consumption of scientific evidence. *Michigan Law Review*, **97**, 931–984.

DROR, I. E. (2005) Technology and human expertise: some do's and don'ts. *Biometric Technology Today*, **13**(9), 7–9.

DROR, I. E. (2006a) Cognitive science serving security: assuring useable and efficient biometric and technological solutions. *Aviation Security International*, **12**(3), 21–28.

DROR, I. E. (2006b) A holistic-cognitive approach for success in technology. *Biometric Technology Today*, **14**(8), 7–8.

DROR, I. E. (Ed.) (2007a) *Cognitive Technologies and the Pragmatics of Cognition*. John Benjamin Press: Amsterdam, 186 pp.

DROR, I. E. (2007b) Land mines and gold mines in cognitive technologies. In I. E. Dror (Ed.) *Cognitive Technologies and the Pragmatics of Cognition*. John Benjamins Publishing: Amsterdam.

DROR, I. E. (in press). How can Francis Bacon help forensic science? The four idols of human biases. *Jurimetrics: The Journal of Law, Science, and Technology.*

DROR, I. E. & CHARLTON, D. (2006) Why experts make errors. *Journal of Forensic Identification*, **56**(4), 600–616.

DROR, I. E., CHARLTON, D. & PERON, A. (2006) Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, **156**(1), 74–78.

DROR, I. E. & HARNAD, S. (EDS.) (2008a) *Cognition Distributed: How Cognitive Technology Extends Our Minds*. John Benjamins: Amsterdam, 258 pp.

DROR, I. E. & HARNAD, S. (2008b) Offloading cognition onto cognitive technology. In I. Dror & S. Harnad (Eds.) *Cognition Distributed: How Cognitive Technology Extends Our Minds*. John Benjamins Publishing: Amsterdam.

DROR, I. E., PERON, A., HIND, S. & CHARLTON, D. (2005) When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology*, **19**(6), 799–809.

DROR, I. E. & ROSENTHAL, R. (2008) Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences*, **53**(4), 900–903.

DROR, I. E. & SHAIKH, A. (2005a) Face recognition technology: cognitive considerations in system design. *United Kingdom Passport Services (UKPS) Technical Report*.

DROR, I. E. & SHAIKH, A. (2005b) Training for expertise in face recognition and working with face recognition technology (TNA). *United Kingdom Passport Services (UKPS) Technical Report*.

HABER, L. & HABER, R. N. (2004) Error rates for human latent fingerprint examiners. In N. Ratha & R. Bolle (Eds.) *Automatic Fingerprint Recognition Systems*. Springer: New York.

HENRY, E. R. (1900) *Classification and Uses of Fingerprints*. Routledge and Sons: London.

HUGHES, T. P. (1983) *Networks of Power*. Johns Hopkins University Press: Baltimore, MD.

INDOVINO, M. *et al.* (2009) *ELFT Phase II—An Evaluation of Automated Latent Fingerprint Identification Technologies*, NISTIR 7577, available at http://fingerprint.nist.gov/latent/NISTIR_7577_ELFT_PhaseII.pdf.

KAYE, D. H. (2009) Rounding up the usual suspects: a logical and legal analysis of DNA trawling cases. *North Carolina Law Review*, **87**, 425–503.

KLUG, D., PETERSON, J. & STONEY, D. (1992) *Automated Fingerprint Identification Systems: Their Acquisition, Management, Performance and Organizational Impact*. National Institute of Justice: Washington DC, NCI 13749.

KOEHLER, J. J. (1993) The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, **56**, 28–55.

KOMARINSKI, P. (2005) *Automated Fingerprint Identification Systems (AFIS)*. Elsevier: Amsterdam.

KOMARINSKI, P. D. (2009) *Considerations for Improving Latent Print Processing, Slides from presentation for NIST Latent Fingerprint Testing Workshop*, available at http://fingerprint.nist.gov/latent/workshop09/Komarinski.pdf.

MEAGHER, S. (2009) *Defining AFIS Latent Print "Lights-Out", slides from presentation at NIST ELFT Workshop*, available at http://fingerprint.nist.gov/latent/workshop09/DefineLPlightsout.pdf.

MNOOKIN, J. L. (2001) Fingerprint evidence in the age of DNA profiling. *Brooklyn Law Review*, **67**(Fall), 13–70.

MNOOKIN, J. L. (2004) The Achilles' heel of fingerprints. *Washington Post*, A27, May 29, 2004.

MNOOKIN, J. L. (2008a) Of black boxes, instruments, and experts: testing the validity of forensic science. *Episteme*, **5**, 343–357.

MNOOKIN, J. L. (2008b) The validity of latent fingerprint identification: confessions of a fingerprinting moderate. *Law, Probability and Risk*, **7**, 127–141.

MNOOKIN, J. L. (in press). The Ira. M. Belfer Lecture, Brooklyn Law School: the future of forensic science. *Brooklyn Law Review*, **75** (in press).

NATIONAL RESEARCH COUNCIL. (1996) *The Evaluation of Forensic DNA Evidence*. National Academy Press: Washington, DC.

NEUMANN, C., CHAMPOD, C., PUCH-SOLIS, R., EGLI, N., ANTHONIOZ, A. & BROMAGE-GRIFFITHS, A. (2007) Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Science*, **52**(1), 54–64.

OFFICE OF THE INSPECTOR GENERAL, OVERSIGHT AND REVIEW DIVISION, U.S. DEPARTMENT OF JUSTICE. (2006) *A Review of the FBI's Handling of the Brandon Mayfield Case*. http://www.justice.gov/oig/special/s0601/PDF_list.htm.

People v. *Collins*, (1968) 438 P. 2d 33.

PETERSON, J. L. & MOORE, J. (1996) The status of AFIS systems worldwide: issues of organization, performance and impact. In Joseph Almog & Eliot Springer (Eds.) *Proceedings of the International Symposium on Fingerprint Detection and Identification*. Israel National Police: Ne'urim, Israel.

POLSON, C. J. (1950) Finger prints and finger printing: an historical study. *Journal of Criminal Law and Criminology*, **41**, 495–517.

POLSON, C. J. (1951) Finger prints and finger printing. An historical study (concluded). *Journal of Criminal Law and Criminology*, **41**, 690–704.

RISINGER, D. M. *et al.* (2002) The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation an suggestion. *California Law Review*, **90**, 1–56.

SCHUM, D. A. (2001) *The Evidential Foundations of Probabilistic Reasoning*. Northwestern University Press.

SHAFFER, J. P. (1995) Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561–584.

STACEY, R. B. (2004) Report on the erroneous fingerprint individualization bombing case. *Journal of Forensic Identification*, **54**(6), 706–718.

STONEY, D. A. (1991) What made us ever think we could individualize using statistics. *Journal of the Forensic Science Society*, **31**(2), 197–199.

United States v. *Mitchell* (2004) 365 F. 3d 215 (3d. Cir.)

WEBER, M. (1949) *The Methodology of the Social Sciences* (Edward A. Shils & Henry A. Finch, Trans. & Eds.).

ZABELL, S. (2005) Fingerprint evidence. *Journal of Law & Policy*, **13**, 143–179.