A FLOATING CONJECTURE:
IDENTIFICATION THROUGH FACIAL RECOGNITION

Wade Robison

Rochester Institute of Technology (United States)

wlrgsh@rit.edu

ABSTRACT

Using facial recognition as evidence in trials is part of a larger pattern of using suspect marks of identification to pinpoint those responsible for crimes. Those accused of crimes have been convicted on the basis of bite marks, hair samples, and fingerprints, and though those in law enforcement would no doubt want a mode of identification that ensures that those accused are in fact those who committed the crime being prosecuted, facial recognition technology fails to add any certainty to the modes of identification that are unfortunately now used and fail to sort out the guilty from the innocent.

We might well think that technological advances will eventually allow us to find us a sure proof mode of identification, but facial recognition technology is nowhere near the level of certainty of identification that we need to prove someone guilty beyond a reasonable doubt and there are serious doubts that it ever will be.

KEYWORDS: facial recognition, features comparisons, fingerprints, prosecution.

1. INTRODUCTION

Facial recognition technology is now being used by law enforcement and by prosecutors to identify and help convict criminal suspects. It is standard now whenever there is a crime to look at what the security cameras captured. They seem to be ubiquitous, and it is a rare article or broadcast on a criminal investigation in the United States that does not give a nod to how helpful a security camera was to identifying a suspect and, indeed, recording the criminal act. There is little doubt that they have been a huge help for law enforcement and that it is likely that we will see more and more cameras deployed throughout our city's streets, in businesses, and in homes.

That technology combined with the tracking information available via cell phones creates an alarming capacity on the part of governments to know exactly where someone is and what they are doing—and the subsequent concern about constant surveillance and control. That concern will no doubt take a back seat to the demands of law enforcement for its use in identifying possible perpetrators. It is too useful for those in law enforcement to be persuaded that a potential alarming capacity should curtail the deployment of surveillance cameras. Such cameras promise to be even more useful as the technology evolves. So deployment is going to continue, with greater and greater capacities to monitor citizens' activities.

That is disconcerting, but even more disconcerting is the use of facial recognition as evidence by prosecutors. Its use in criminal trials in the United States is part of a larger pattern of using questionable rules of skill that tell us how to identify a suspect based on fingerprints, bite marks, and other supposedly identifying information.

We will look at the rules of skill that are at issue in what is called features comparisons before examining how such features are used, and with what success, in prosecuting cases. We will then turn to facial recognition to examine its advantages and disadvantages in prosecution.

2. RULES OF SKILL

A rule of skill tells us how to achieve a particular end: to bake a cake, do such-and-such; to buttress a girder, do so-and-so. They are the tools of the trade, so to speak, for any profession, and they display a wide variety of functions. There are rules that tell us what things are—what symptoms go with which disease, what crosshair signatures are, what shape and position goes with which human organ, and on and on. There are rules that tell us how to do something—how to extract a tooth, how to use a Japanese saw, how to use Matlab, and on and on. There are rules that prescribe the procedures to follow—in writing a valid will or ensuring a fair trial, in minimizing the risks of infection, and on and on.

We are all familiar with rules of skill. We learn them early on as we learn to count or correct our pronunciation so we can be understood. We know as well what happens when we fail to follow the relevant rules. We open ourselves to criticism and to failure. We learn early on that games must be played in certain ways and not others, for instance, and we learn as well the limits of rules in ensuring that individuals do what they ought to do. Cakes do not always turn out the way they should. Girders are sometimes not properly buttressed. Surgeons amputate the wrong limb. Police officers stop a vehicle merely because the driver is African-American (Wang, 2017).

Rules of skill set a sequenced, coherent normative order to what we do. What matters for making fudge, for instance, is that add vanilla after we have melted in the chocolate for fudge, not before, and that we pay attention only to what is required for making fudge. Scratching one’s head while thinking about where the chocolate might be may occur while making fudge, but it not part of what it is to make fudge. What is required to make fudge is a coherent series of steps, and anything else that may occur while traversing those steps is not relevant. That is why the rule is normative: it tells us what we ought to do to make fudge and, in doing that, tells us what we ought to ignore as not part of the sequenced, coherent, normative order

Rules of skill are no different in that way than, say, the rules of logic. Valid argument forms, for instance, tell us how we ought to reason deductively. When we provide someone with the form for modus ponens—if p, then q and p, therefore q—we are providing them with a sequence of steps that they ought to follow if they are not to risk reasoning from truth to falsity. The same is true for any rule of calculation. That is why tellers at the grocery store cannot just give us any old handful of change. They are constrained by the rules that tell us how we ought to add and subtract. If they fail to give us the correct change, we may properly tell them that they have made a mistake, done something they ought not to do. We are telling them, effectively, that they are not doing what reason tells all of us we ought to do. When we subtract 76 cents from a dollar, we do not, with good reason, get 21 cents. The norm we have failed to satisfy is one of reason

3. FLOATING CONJECTURE

Some rules of skill are floating conjectures, without the sorts of evidential backing needed to make them reliable. Anyone who reads mysteries or watches crime shows knows that central to a crime's solution is what can be found at the crime scene—'DNA, hair, latent fingerprints, firearms and spent ammunition, toolmarks and bitemarks, shoeprints and tire tracks, and handwriting' (Report, 2016). Detectives hunt for samples at the scene that can then be compared to samples from a suspect, and they remind everyone not to touch anything at the scene so that when they dust for fingerprints, for instance, their findings will not have been contaminated. They are hunting for fingerprints or hair or something else left by whoever committed the scene crime.

Experts compare the features of what is found at the crime scene with the features of the relevant sample from a suspect, and if there is a match, they have significant evidence that the suspect is the criminal. The relevant rule of skill will vary depending upon what feature is being examined, but the general formula is the same for all features: if this sample looks like that sample, they are from the same person.

We can already see, from the vagueness of that formulation, how easy it must be for errors to enter into any identification. 'Looks like' requires someone to do the looking, and so one source of error is that the person doing the looking may make mistakes. Another source of error concerns in what way or ways the items being inspected look alike—and unalike. My siblings and I have a family resemblance and so look alike in certain ways, but not in others. What feature or features should count or count the most—the shape of our ears, their position relevant to our skulls, our noses, the shape of our nostrils, our projecting or receding chins, or what? What is compared with what requires a judgment based on evidence of which features, if any, are telling. And so, clearly, another potential source of error is the misidentification of what ought to count when comparing samples, and then yet another is the judgment that two samples are identical in regard to what has been judged to be the telling feature.

We can get a sense of how problematic the relevant rules of skill are by examining the track record of identifications for bite marks and hairs and fingerprints. We have a standard we can use in DNA.

DNA analysts don't tell jurors that a suspect is a match. Instead, they use percentages. Because we know the frequency with which specific DNA markers are distributed across the population, analysts can calculate the odds that anyone other than the suspect was the source of the DNA in question (Balko, 2020).

We have a basis for comparison with DNA. Since we know of any specific DNA marker how many there are in the population, we can tell how probable it is that one DNA marker is like another.

But we should emphasize that DNA is not the gold standard we may think it is. For one thing, it is not foolproof. A man who received a bone marrow transplant ended up with the DNA of the donor. He had both his own and his donor's DNA, and, somewhat to his chagrin, we must assume, 'all of the DNA in his semen belonged to his donor' (Murphy, 2019). We do not have any idea how often that happens, but once is enough to make DNA testing less than the gold standard.

There are also problems with how the testing is done. Those doing it can make mistakes, obviously, sending the police on the sort of fool's errand the German police engaged in for sixteen years after finding 'traces of identical female DNA…at 40 crime scenes across southern

Germany and Austria,' including six murders (DNA, 2009). The police used Q-tips that they had purchased from a store rather than sanitized ones, and the Q-tips had been contaminated by a woman working at a Q-tip factory in Bavaria.

So using DNA to tie a particular suspect to a crime scene is not without its problems (Otterman, 2019). It is, however, determinative enough that we can assess the validity and reliability of comparing hair and bite marks, for example, by determining if using DNA gives us the results we got in previous cases comparing other features from the crime scene.

4. HAIR AND BITE MARK

Santae Tribble was 17 when he was arrested for murder and convicted based on a comparison of his hair with hair found at the scene of the crime. As he put it, the experts said that the sample 'matched my hair in all microscope characteristics.' As the prosecutor said in summing up the evidence,'There is one chance…in 10 million that it could [be] someone else's hair' (Hsu, 2012a). Later analysis showed that of the thirteen hairs in question, nine were from one person, three from different individuals, and one from a dog. None belonged to Tribble (Oliver, 2017). He was freed (Hsu, 2012b) and then exonerated (Hsu, 2012c), but he spent 26 years in jail because of the mistaken judgment that his hair matched the samples found at the crime scene.

'Such is the true state of hair microscopy,' the lawyer representing Tribble said, that '[t]wo FBI-trained analysts, James Hilverda and Harold Deadman, could not even distinguish human hairs from canine hairs.' Researchers showed in 1974 that 'visual comparisons are so subjective that different analysts can reach different conclusions about the same hair. The FBI acknowledged in 1984 that such analysis cannot positively determine that a hair found at a crime scene belongs to one particular person' (Hsu, 2012a).

In 2012, the FBI and Department of Justice began a review of over 3000 'criminal cases involving microscopic hair analysis.' They found that 'that FBI examiners had provided scientifically invalid testimony in more than 95 percent of cases where that testimony was used to inculpate a defendant at trial' (Report 2016). So 19 out of every 20 defendants were falsely incriminated by FBI experts. It is difficult to imagine a less reliable way to determine if someone has committed a crime. Flipping a coin would give better than a 95% failure rate

Another example of a floating conjecture in forensic science concerns bite marks. Keith Harward 'narrowly escaped the death penalty,' but spent 33 years in prison after being convicted of rape and murder on the basis of six forensic dentists testifying that the bite marks on the rape victim's legs were his. DNA evidence showed that he was innocent and that a fellow sailor, Jerry Crotty, was responsible. Harward is one of at least 25 individuals 'to have been wrongfully convicted or indicted based at least in part on bite mark evidence' (Innocence, 2019). He is now free, but he says to those who tell him he is a free man, 'I will never be free of this…I spent more than half my life in prison behind the opinions and expert egos of two odontologists

Harward noted that there was 'a death-penalty case in Pennsylvania where the judge is going to allow bite-mark evidence' (Oliver, 2017). Indeed, 'bite-mark analysis…has yet to be disallowed by any courtroom in the country' (Balko, 2020)

The 2016 Report to the President pointed out that a '2010 study of experimentally created bitemarks…found that skin deformation distorts bitemarks so substantially and so variably that current procedures for comparing bitemarks are unable to reliably exclude or include a suspect as a potential biter.' In fact, evidence 'showed a disturbing lack of consistency in the way that forensic odontologists go about analyzing bitemarks, including even on deciding whether there was sufficient evidence to determine whether a photographed bitemark was a human bitemark' (Report, 2016). That bite mark evidence still finds its way into court cases is a sad commentary on the failure of American judicial system to come to grips with such forensic floaters.

5. FINGERPRINT

On March 11, 2004, ten bombs killed 192 passengers on trains in Madrid and injured more than 1400, according to initial reports (Sciolino, 2004). The Spanish authorities found a fingerprint on a bag of detonators and forwarded it to the FBI to see if it could find a match in its database. The FBI's Integrated Automated Fingerprint Identification System (IAFIS) 'generated a list of 20 candidate prints.' None was a perfect match, but IAFIS also lists close matches, and one belonged to Brandon Mayfield, a lawyer in Oregon. The FBI 'immediately opened an intensive investigation of Mayfield, including 24-hour surveillance…and physical searches' of his law office and residence. When news somehow broke that an American was a suspect in the bombing, the FBI detained Mayfield on May 6th because they were 'absolutely confident' that Mayfield's fingerprint was on the detonator bags. They kept him in solitary confinement 'for up to 22 hours per day' (Office, 2006)

The fingerprint from Spain was examined by a fingerprint specialist in the FBI who verified it as belonging to Mayfield. That judgment was confirmed by a second FBI fingerprint specialist and by the fingerprint unit chief, all of whom agreed it was Mayfield's. That decision was confirmed by a court-appointed specialist (Office, 2006). Four fingerprint experts fingered Mayfield, as it were.

The defense attorney's own expert confirmed the judgment of the FBI experts and later said, 'No time before in history have there ever been two fingerprints with fifteen minutiae that were not the same person' (Bharara, 2020). So there was good reason for the FBI's confidence.

The Spanish authorities identified the person whose fingerprint was on the bag of detonators, and it was not Mayfield. As it turned out, further analysis of the fingerprints showed that Mayfield's was not identical to the one found in Spain, but what is of importance here is that specialists in fingerprint identification judged that it was and that they had absolute confidence in their judgment. The Mayfield case is a dramatic example of why such judgments cannot be relied upon and should not be relied on, especially in criminal cases where the stakes are high. We must have proof beyond a reasonable doubt, and the Mayfield case puts in doubt reliance on fingerprints comparisons. The case has become a classic example of how misidentification of a sample can mislead investigators, taking them off the scent of the perpetrator onto the scent of an innocent person who can be badly harmed by the mistake.

6. RELIABILITY

As it turns out, feature comparisons are not very reliable at all. The 2016 Report to the President on forensic science stated,

Reviews by the National Institute of Justice and others have found that DNA testing during the course of investigations has cleared tens of thousands of suspects and that DNA-based re-examination of past cases has led so far to the exonerations of 342 defendants (Report, 2016).

The failure of such feature comparisons as hair samples and fingerprints is illustrated by the number of exonerations each year as old cases are reexamined. 'More than 150 men and women were exonerated in 2018,' having 'spent more than 1,600 years in prison' for crimes they did not commit. The Innocence Project exonerated more than 350 individuals, and in 45% of the cases, those individuals were convicted because of a failure of feature comparisons combined with misleading testimony from experts who ensured juries and judges that they were sure within a 'reasonable degree of scientific certainty.' But the 'experts…used exaggerated statistical claims to bolster unscientific assertions.' That is a phrase that a jury is likely to believe gives great weight to the evidence but has no scientific validity (Innocence, 2019).

The 2016 Report quotes a judge about testimony from an expert that 'markings on certain bullets were unique to a gun recovered from a defendant's apartment':

As matters currently stand, a certainty statement regarding toolmark pattern matching has the same probative value as the vision of a psychic: it reflects nothing more than the individual's foundationless faith in what he believes to be true. This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—real proof—beyond a reasonable doubt, precisely because the stakes are so high.

The Report adds,

> In science, assertions that a metrological method is more accurate than has been empirically demonstrated are rightly regarded as mere speculation, not valid conclusions that merit credence (Report, 2016).
>
> The need for evidence and testimony based on evidence is nicely put by U.S. District Judge John Potter, in 'an early case on the use of DNA analysis,' U.S. v. Yee (1991):
>
> Without the probability assessment, the jury does not know what to make of the fact that the patterns match: the jury does not know whether the patterns are as common as pictures with two eyes, or as unique as the Mona Lisa (Report, 2016).

That, in a nutshell, is the problem with the comparison of features: 'There is no way to calculate a margin for error.' Unlike DNA testing, where we know how probable it is that one marker is like another because we know of any specific DNA marker how many there are in the population, comparing a hair found at the scene of a crime to one of a suspect can at best exclude it—if, say, the one is blond and other one black. Depending on how many features of a hair sample are compared, it may not exclude many at all.

'[T]he FBI agent testified at trial that the hair from the stocking matched Tribble's "in all microscopic characteristics",' (Hsu, 2012c), but the FBI expert, Hilverda, 'recorded in his lab notes that he had measured only three characteristics of the hair…—it was black, it was a human head hair, and it was from an African American' (Hsu, 2012a). We can presume that under a microscope more than three characteristics are discernible and that countless African Americans have black hair. So the FBI agent's testimony was misleading, to say the least, and Tribble spent 26 years in jail for being an African American.

Here we know that the three characteristics are hardly unique, but no matter how many features are found, we would have no idea how many hairs in the world share those features. The hairs may even match in all discernible ways, but with no idea how many different hairs of different individuals match, we have no idea whether the match is unique or only to be expected since millions could match.

The same is true for any comparison. We cannot know we have a unique match with marks on shell casings, or bite marks, or pry marks on a door because there is no way of knowing how many different guns or teeth or crowbars might, under the right conditions, produce identical marks (Balko, 2020).

We have floaters in forensic science. Because of them, some individuals were executed. Floaters can have grievous consequences, and those professionals who testified to their valid application in particular cases were wrong.

7. DEGREES OF CONFIDENC

It is no surprise that people can be confident about something or find something plausible or even obvious when the facts do not warrant confidence. We all have beliefs which range from the implausible to certain, and to assess them, we must rely not on how we feel about them, but on what the facts support. A feeling that a belief is certain is no guarantee it is true. If we were to construct an argument for the FBI experts' judgment that Tribble's hair was found at the scene of the crime, it would include the following premises regarding the degree of confidence the FBI experts had in their judgment implicating Tribble:

- There is one chance in 10 million that the hair is not Tribble's.
- We have only been mistaken 19 times out of 20 in making such judgments.
- So we experts are absolutely confident the hair is Tribble's.

We have terms of criticism for beliefs, and the one most relevant here concerns the degree of likelihood that the belief is true. We ought to be more or less confident in our beliefs in accordance with the quality of our evidence, and in this case we ought to lack any confidence at all.

A birder trying to identify a particular warbler will follow the usual methodology, making a judgment based on the bird's size, flight pattern, song, and other distinctive characteristics. The birder ought to be more or less confident depending upon how many identifying marks are discernible and how easily they can be discerned. 'It's a Palm Warbler' is a quite different judgment than 'Well, could be a Palm Warbler,' and they mark how many identifying marks the birder was able to discern and with what degree of certainty. Does the bird have a distinctive yellow eyebrow? A chestnut-colored crown (Sibley, 2000)? Catching a glimpse of something chestnut-colored is very different from being able to observe the bird for some period of time.

The methodology for identifying birds is not perfect. Experts can use it and still make mistakes. But when used correctly by a competent birder, the success rate is significantly higher than 5%. A 95% failure rate tells us that the methodology is unreliable and that having a second and third expert check another's judgment using the same methodology will not provide us with any more evidence for the truth of the belief.

If the methodology is faulty, it does not matter how experienced an expert may be, or how many experts chime in. An unreliable methodology will lead to unreliable results. As the President's Report of 2016 put it, Without appropriate estimates of accuracy [and error rates], an examiner's statement that two samples are similar–or even indistinguishable–is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact.

As the Report notes,

> Nothing–not training, personal experience, nor professional practices–can substitute for adequate empirical demonstration of accuracy (Report, 2016).

The rules of skill that supposedly gave credence to 'expert' testimony are all recipes for mistakes. In comparing Mayfield's fingerprint with the fingerprint from the bag of detonators, FBI fingerprint specialists found ten points of similarity, and the defense's expert found fifteen. 'Points' is a technical term here. They occur where individual ridges end or split, and the similarities were 'the relative location of the points, the orientation of the ridges coming into the points, and the number of intervening ridges between the points.' The Office of Inspector General's Review of the FBI's Handling of the Brandon Mayfield Case points out that there is no research on how frequently such similar constellations of points occur in different individuals, but that 'anecdotal reports suggest that this degree of similarity…is an extremely unusual circumstance' (Office, 2019).

The bottom line, however, is that the experts were relying on a rule of skill that told them that if there are so many points of comparison between two fingerprints, they can have 'absolute confidence' that the two were made by the same person when they have no way to gauge a margin of error. We have no idea how often such a constellation of points occurs among all the fingers in the world, and without that information, we can only use a particular constellation of points as a way of excluding some possible suspects. We cannot pinpoint a suspect because we have no idea how many others share the relevant constellation. So the rule of skill the experts used is a floater, a recipe that provides no justification for any confidence at all in its outcome.

The other floaters are no better supported. They all depend on rules of skill that tell supposed experts that if they have such-and-such a configuration in two samples—of markings on a bullet, of the impression of teeth marks, of the details of hair—they can be absolutely confident that the samples came from the same firearm, or the same mouth, or the same head of hair. Such confidence is not responsive to reality, but reflects an unwarranted judgment about the reliability of a faulty rule of skill (National Research Council, 2009)

It is not just experts who make mistaken judgments about feature similarities. Eye-witness identifications are standard and are remarkably unreliable. A witness or the victim to a mugging gets a glance at someone's face and then identifies the defendant when asked by a prosecutor in the courtroom to point out the person responsible for the crime, but 'inaccurate eyewitness identifications…were introduced as evidence in over 70 percent of the more than 360 cases that the Innocence Project… proved were wrongful convictions' (Rakoff, 2019). Amateurs are no better than experts, that is.

8. FACIAL RECOGNITION

The use of facial recognition technology only adds to the floaters, with additional problems. The history of floating conjectures—fingerprints, bite marks, and so on—brings out most of the problems that plague using facial recognition for identifying and prosecuting suspects

The main problems are the ones that plagued the FBI when it misidentified Brandon Mayfield. His fingerprint was not a perfect match, but one of twenty that the FBI algorithm picked out from its massive data base as closely similar. The algorithm did not get a direct hit, but twenty possibilities, that is, and the failure of the algorithm to provide an exact match meant that judgment calls were necessary to determine which of the twenty, if any, was the most likely match. But without knowing how many individuals have fingerprints that fall within that range of possibilities, there is no way of knowing that any one individual has been properly identified. The most that can be said is that those individuals whose fingerprints are not within that range are not suspects.

This summary of the main problems captures the essence of what is wrong with using facial identification: the failure to zero in on exactly one person by comparing features requires judgment calls with no way of knowing the likelihood of one's getting it right. But this summary also obscures just how difficult a facial features judgment is.

We know that for fingerprints the likelihood of finding an exact match is exceedingly small. No matter how detailed the FBI's sample may be, it is being compared to a sample from a crime scene. It is highly unlikely that the two prints are going to match exactly. Smudging, a lighter touch here and a heavier touch there, a twist as one lets go, degradation through exposure to contaminants—all sorts of things can get in the way of a clear print. The consequence is that there are always gaps that need to be filled, and where there are gaps requiring judgment calls on the part of those doing the examination, we have an opening, a gap, that is, for mistakes

Facial identification also faces that issue, so to speak. No two facial images of a person are any more likely to be identical in all respects than two fingerprints of a person. What feature or features should count or count the most? That is the first decision to be made, and it is not at all clear what to choose to minimize mis-identification. A head turned slightly away, the beginnings of a smile, an irritated expression, a new hairstyle that covers, or uncovers, parts of one's face—all sorts of things can get in the way of a perfect match

What counts cannot be the whole face, presumably, because any two images are almost certainly going to vary because of the angles from which they are taken or the direction a person is facing or the quality of the image itself. What counts cannot be anything that changes when one's facial expression changes. Just imagine how different a person's face can look when the person is smiling, frowning, angry, grinning from ear to ear, disgusted, sad, pouting, eye-rolling, surprised, and so on.

Whatever is chosen as the telling feature or features must be constant, invariable despite different camera angles, different positions of the person's head, or different emotional expressions. What is telling must not alter no matter what other differences there are. Human faces do share some relatively constant characteristics. If you want to draw a face, you need to start with an oval, divide it vertically and horizontally in half, place the eyes on the horizontal line on each side of the vertical line, and so on with the nose and lips and ears and other features taking their usual places. But although those features are relatively constant, they are not always so, and in any event, they are too general to allow anyone to zero in on any one face using those constants. It is unclear what other feature or features would provide the detail and constancy needed to make an identification. It is unclear, that is, how we are to complete the first step of determining what is telling.

In any event, however that first determination is made, we have another problem. Whatever the telling feature may be, it is not going to distinguish between identical twins or, presumably, doppelgängers. Identical twins look alike, obviously, and doppelgängers look enough alike that, in the best—or the worst of cases for an investigator or prosecutor—they may as well be identical twins. We know of cases where identical twins were separated at birth, neither knowing of the other, only to discover one another years later (Paparella et al., 2018). What we do not know is how many identical twins have been separated and have never found out that they were identical twins. The number affects how likely it is that an identification based on facial recognition is correct, but without that information, we cannot be at all sure what the likelihood is that we have a match.

And there are more than a few individuals who look alike without being so much alike we would call them doppelgängers. I have been mistaken countless times for the actor in the Halloween series (sans costume), once by a flight attendant who said, 'Oh!' when she saw me, asked why I was sitting in the very back of the plane, told me she would make sure I was not disturbed when I said, 'For peace and quiet,' and then, as I left, having presumably looking up my name, commiserated with me for not being famous. I was not quick thinking enough to tell her that I always travel under my real name

So even if a determination can be made of what counts as a tell, we are no better off than we were with the other floating conjectures. We can only exclude some and not pinpoint any particular person. That problem ought to be sufficient to rule out facial recognition as definitive to prove guilt beyond a shadow of a doubt. As we quoted above on the reliability of bullet markings, 'This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—real proof—beyond a reasonable doubt, precisely because the stakes are so high.

But facial recognition faces other problems. One I have not mentioned before, but will occur regardless of what features are being compared—police misconduct. With Photoshop and other editing software, it is easy to manipulate images to fill in the gaps that will occur when comparing one image with another. We already have evidence that police investigators have doctored photos to increase the likelihood of a hit and so the chances of an arrest. 'Some investigators,' it has been reported, 'edited the photos in hopes of revealing more matches, including swapping out facial features, blurring or combining parts of photos and pasting in images of other people's lips or eyes' (Harwell, 2019a)

Facial recognition software faces yet another problem. It is biassed. The algorithms used misidentify asians and blacks far more often than whites, females far more often than males, and native Americans most of all. 'Middle-aged white men generally benefitted from the highest accuracy rates.'

The National Institute of Standards and Technology, the federal laboratory known as NIST that develops standards for new technology, found 'empirical evidence' that most of the facial-recognition algorithms exhibit 'demographic differentials' that can worsen their accuracy based on a person's age, gender or race (Harwell, 2019b).

We have a situation rather like the one we faced when airbags were first introduced. The engineers chose as their norm, the one best protected by the exploding airbag, five-foot-nine males weighing 170 pounds, a choice that best protected the 50th percentile of men and the 95th percentile of women. It is not a far reach to wonder whether the sex of those designing the airbags mattered (Robison, 2016)

Just so, it is not a far reach to wonder about the race and sex of those designing facial recognition software. The problem is similar to the soap dispenser that fails to recognize any hands but those of whites (Hale, 2017). Clearly those who designed it failed to test it across the range of diverse hands that it would need to recognize.

The point is that the algorithms experts use to help fill in the gaps are a function of the vast amount of data now available to be mined for accurate assessments, but that data captures our biasses as well. 'In 2015, for example, the Google Photos app was caught labeling African-Americans as "gorillas"' (Metz, 2019). Such mistakes can be corrected, but that misses the point. They can permeate the algorithms used in facial recognition, and until we find such mistakes, we have no idea how frequently they occur and how they can bias the results. We are in the same position as those who use fingerprints, bite marks, bullet markings and other features. Without knowing how often we will find the same bullet markings in all the guns in the world, the best we can do is to exclude some, leaving however large an unknown number suspect. Just so, because we have no idea how often such mistakes in algorithms occur, the best we can do is exclude some individuals, perhaps leaving an impossibly large number in the pool of suspects.

9. CONCLUSION

The bottom line is that facial recognition cannot be any more definitive in establishing guilt or innocence than fingerprints or any other feature. The rule of skill experts must use regarding facial recognition floats, without the sorts of evidential backing needed to make it reliable. It shares the two failings we identified for the other features being compared:

1. We do not know how many share the particular features that are taken to be telling. We do not know, for instance, how many individuals have fifteen or ten identical points in a fingerprint or how many have such-and-such a distance between the centers of their eyes.
2. The gaps that are found between any two samples must be filled by the judgment of an expert, but without any knowledge of how many individuals share the original configuration, how large a pool there is, that is, no expert, no matter how experienced, can provide a knowledgeable judgment about any particular individual or, indeed, even a judgment of the likelihood of a particular person being the person to be charged or convicted.

So facial recognition is not some new and wonderfully different and effective technique for identifying and prosecuting individuals. The advantage of providing a history of features comparisons and the ways in which they have been helpful and harmful is that we can see that comparing facial features is more of the same, with all the problems we already canvassed and more.

That is not to diminish its value. Law enforcement can and no doubt will use it to rule out a relatively large class of individuals in any specific case. The images are good enough to allow for a great deal of discrimination, and that can be crucial to those trying to track down a suspect. But it will have no value for prosecutors. A supposed match would only show that a defendant is one of an indeterminate number of individuals with relevantly similar features.

REFERENCES

Balko, Radley. (2020, February 28). A D.C. judge issues a much-needed opinion on “junk science.” The Washington Post.

Bharara, Preet. (2020). Doing Justice. New York: Vintage Books.

'DNA bungle' haunts German police. Retrieved from http://news.bbc.co.uk/2/hi/europe/7966641.stm.

Hale, Tom. (2017, August 17). This Viral Video Of A Racist Soap Dispenser Reveals A Much, Much Bigger Problem. IFLScience. Retrieved from https://www.iflscience.com/technology/this-racist-soap-dispenser-reveals-why-diversity-in-tech-is-muchneeded/.

Harwell, Drew (2019a, May 16). Police have used celebrity lookalikes, distorted images to boost facial-recognition results, research finds. The Washington Post.

Harwell, Drew (2019b, December 19). Federal study confirms racial bias of many facial recognition systems, casts doubt on their expanding use. The Washington Post.

Hsu, Spencer S. (2012a, April 16). Convicted defendants left uninformed of forensic flaws found by Justice Dept. The Washington Post.

Harwell, Drew (2012b, May 16). Santae Tribble’s 1980 murder conviction overturned by D.C. judge. The Washington Post.

Harwell, Drew (2012c, December 14). Santae Tribble cleared in 1978 murder based on DNA hair test. The Washington Post.

The Innocence Project: Keith Allen Harward (2019). Retrieved from https://www.innocenceproject.org/cases/keith-allen-harward/.

Metz, Cade (2019, November 11). We Teach A.I. Systems Everything, Including Our Biases. The New York Times.

Murphy, Heather. (2019, December 7). When a DNA Test Says You’re a Younger Man, Who Lives 5,000 Miles Away. New York Times.

National Research Council of the National Academies. (2009). Strengthening Forensic Science in the United States: A Path Forward. Washington D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/12589.html

Office of Inspector General (2006). A Review of the FBI’s Handling of the Brandon Mayfield Case. Retrieved from https://oig.justice.gov/special/s0601/exec.pdf.

Oliver, John (2017, October 1. Forensic Science: Last Week Tonight. Retrieved from https://www.youtube.com/watch?v=ScmJvmzDcG0; from 13:21 to 14:40.

Otterman, Sharon (2019, April 23). She Was Fired After Raising Questions About a DNA Test. Now She’s Getting $1 Million. The New York Times.

Paparella, Andrew et al. (2018, March 9). Twins make astonishing discovery that they were separated shortly after birth and then part of a secret study. Retrieved from https://abcnews.go.com/US/twins-make-astonishing-discovery-separated-birth-part-secret/story?id=53593943.

Rakoff, Jed S. (2019, April 18). Our Lying Eyes. New York Review of Books.

Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity in Feature-Comparison Methods. (2016). Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Robison, Wade. (2016). Ethics Within Engineering, An Introduction. London: Bloomsbury Academic Publishing.

Sciolino, Elaine. (2004, March 12). BOMBINGS IN MADRID: THE ATTACK: 10 Bombs Shatter Trains in Madrid, Killing 192. The New York Times.

Sibley, David Allen. (2000). The Sibley Guide to Birds. New York: Alfred A. Knopf.

Wang, Amy B. (2017, July 13). Video shows police trying to explain why they pulled over a Florida state attorney. The Washington Post.