# Is Forensic Science in Crisis?

Michał Sikorski

April 12, 2022

**Abstract**

The results of forensic science are believed to be reliable, and are widely used in support of verdicts around the world. However, due to the lack of suitable empirical studies, we actually know very little about the reliability of such results. In this paper, I argue that phenomena analogous to the main culprits for the replication crisis in psychology (questionable research practices, publication bias, or funding bias) are also present in forensic science. Therefore forensic results are significantly less reliable than is commonly believed. I conclude that in order to obtain reliable estimates for the reliability of forensic results, we need to conduct studies analogous to the large-scale replication projects in psychology. Additionally, I point to some ways for improving the reliability of forensic science, inspired by the reforms proposed in response to the Replicability Crisis.

**Keywords:** Forensic Science; Reliability; Replicability Crisis; Questionable research practices; Funding bias; Wrongful convictions

## 1 Introduction

Forensic science is a collection of disciplines that utilize scientific methodologies in criminal (and less often civil) investigation. The forensic disciplines include DNA analysis, fingerprint analysis, and microscopic hair analysis. Because of the employed methodology, forensic procedures and their results are expected to be as reliable as those of the natural sciences. Consequently, forensic results are widely believed to be reliable by judges, jurors (see e.g., Koehler 2016b) and experts (see e.g., Murrie et al. 2019), and are widely used in support of verdicts around the world. However, research indicates that, due to the lack of suitable empirical studies, we actually know very little about the reliability of such results. Additionally, in the last decade, it became clear that forensic science is troubled by a number of methodological problems, and in effect, regularly delivers misleading or even false results which in turn lead to wrongful convictions (see e.g., NRC 2009).

In this paper, I will attempt to re-assess the reliability of forensic results. I will rely on two sources. First, my conclusions and recommendations are inspired by the rich methodological discussion surrounding the replication crisis in scientific disciplines such as psychology. Second, my

review of the reliability of forensic science will be partly based on two recent reports: (NRC 2009) and (PCAST 2016), that examined the state of forensic science by reviewing existing evidence. The reports were presented by: the National Research Council and the President's Council of Advisors on Science and Technology in response to the emerging doubts concerning the reliability of forensic results.

The analogy between the replication crisis and the state of forensic science has been drawn in the literature (see e.g., Chin 2014 or Chin, Ribeiro, and Rairden 2019), but has not yet been fully explored. There are many promising proposals to rectify the replication crisis. There is even evidence to suggest that we have achieved some progress in improving the reliability of psychology. I argue that the main causes of the replication crisis are also present in forensic science. Therefore, the reliability of forensic results should be on par to that of psychological results, which are much less reliable than it is commonly assumed. Furthermore, since the two crises are analogous, we can apply the lessons from the discussion on the reliability of psychology to study and improve the reliability of forensic science.

In the second section, I will briefly describe the replicability crisis in psychology. In the third and fourth sections, I will discuss two of the most popular forensic practices (ACE-V fingerprints analysis and Short Tandem Repeat Analysis) and what is known about the reliability of forensic science and the possible presence of corrupting factors. I will focus on these two disciplines because they are both the most reliable and most widely used. In the fifth section, I will argue that there is no clear reason to think that forensic science is less susceptible to the methodological problems that psychology faces. In the last section, I will discuss the most efficient way to test the reliability of forensic science and a few ways to improve it.

## 2 Replicability Crisis in (Psychological) Science

Since the beginning of the twenty-first century, scientists, especially psychologists, have dramatically changed how they view their disciplines. In the words of Pashler and Wagenmakers (2012):

> "Is there currently a crisis of confidence in psychological science reflecting an unprecedented level of doubt among practitioners about the reliability of research findings in the field? It would certainly appear that there is."

This general skepticism (see also Baker 2016) is a reaction to *The Replicability Crisis*, which plagues psychology. The crisis consists in the fact that most of the psychological (or even scientific) results are not replicable. Roughly, an empirical result is replicable if we can reliably reproduce the result by repeating the original empirical study (such as an experiment), otherwise we say that the result isn't replicable. Replicability is considered to be a crucial feature of scientific results (e.g., National Academies of Sciences and Medicine 2019) and the general failure of replicability is considered to be catastrophic by many scientists (e.g., Bishop 2019). At the same time, multiple large-scale replication projects (see e.g., Collaboration 2015 or Camerer et al. 2018) showed that just around 50% of psychological results are replicable.

The severity of the crisis derives from the fact that reliability implies replicability. The reliability of a scientific procedure (e.g., experiment) is the chance that it delivers the right result, positive if the hypothesis is true and negative otherwise. If we assume that the experiments in a given discipline are reliable then, in most cases their results should support the true hypothesis. Therefore, a low replicability rate of experimental results entails low reliability of those results. Consequently, as demonstrated by Ioannidis (2005), in most experimental settings, research claims are more likely to be false than true, and results are often just confirmations of preexisting biases. This is unacceptable.

What are the causes of the crisis and how did it come to light? Scientists have been aware that some features of scientific practice are detrimental to the reliability of its results. The most general cause is a methodological flexibility. At each stage of the scientific process scientists have to make decisions that shape the experimental design and affect the final result (see Wicherts et al. 2016). They can abuse these degrees of freedom by making biased methodological decisions, which increase the probability of a preferred outcome (see Wilholt 2008). Additionally, a scientist can explore many possible methodological decisions, for example, made during the analysis (how to handle outliers, which model to use, etc.). Once the scientist has found a combination of choices that produces the preferred result, she would publish the experiment without mentioning combinations that results in a deviation from the preferred result (see e.g., Simmons, Nelson, and Simonsohn 2011). Similarly, she may decide to continue with collecting evidence (for example by adding participants or extending the length of the study) when the results do not support her preferred conclusion and stop the experiment as soon as she gets the result (see Heide and Grünwald 2017). Those and similar covert manipulations called *Questionable Research Practices* (QRP) are possible because of the flexibility of scientific methodology. The use of QRP increases the probability of the desired result by increasing the probability of false-positives (see e.g., Bakker, Dijk, and Wicherts 2012). Another factor which contributes heavily to the crisis is *Publication Bias* (see e.g., Scargle 2000). It consists in the fact that the publishing policy of most scientific journals heavily favors positive results (which support the existence of the effects in question). Because of that, the published literature does not represent the totality of conducted studies. QRP, publication bias, and the misuse of statistical methodology are considered to be the main causes of the crisis (see e.g., Gigerenzer 2004 or Ioannidis 2005).

The presence of these phenomena has been known for a long time (see e.g., Vandenbroucke 1988 or Swazey, Anderson, and Lewis 1993). Sometimes the resulting unreliability is exemplified in the form of implausible published results (e.g., Bennett et al. 2010 or Bem 2011) or confusing and contradictory literature (see e.g., Bishop 1990). Initially, these results were not treated as evidence of methodological issues. Methodological flexibility was treated as an unavoidable feature of science, the questionable results described as bad apples, and the concerns about replicability were considered to be overblown. However, the disappointing results of large scale replication projects made such responses difficult to defend. In these large scale replication projects, the authors have attempted to replicate a large number of important results from a particular discipline. Because of its scale, such projects can persuasively demonstrate the deficiencies of target experimental setups and indirectly, the examined disciplines. For example, Collaboration (2015) presents the results

of one hundred single replication attempts (each experiment was repeated once) of one hundred results from articles published in the most prominent psychological journals. Less than half of the replication attempts were successful. Such estimates together with implausible results convinced the majority of psychologists that the reliability of psychological results is low.

Several reforms have been proposed in response to the crisis. Improving methodological transparency and promoting preregistration, providing the detailed description of the methodology of the experiment before data are collected, were proposed as countermeasures to QRP (see e.g., Nosek and Lakens 2014 or National Academies of Sciences and Medicine 2018). Various improvements in statistical design were proposed (see e.g., Vazire 2016 or Witte and Zenker 2017). For example, Vazire (2016) proposed to set up a requirement of minimal statistical power to exclude the underpowered and therefore unreliable experiments. Statistical power is, roughly, the probability that the test rejects the null hypothesis if the alternative hypothesis is true. It was shown that low statistical power reduces the reliability of an experiment (see e.g., Ioannidis 2005). It can be increased by including additional participants but that makes the experiment more expensive. In effect, underpowered designs were common in psychology. Fixing the acceptable level of power (0.8 in case of the proposal from Vazire 2016) promotes well-powered designs and therefore more reliable results. Finally, replications were encouraged (see e.g., Nosek and Lakens 2014, Peels 2019 or Romero 2018). The replication studies were, until recently, very hard to publish which discouraged scientists from conducting them and was one of the reasons why poor replicability remained largely hidden for so long.

The most recent meta-scientific results validated both concerns over the reliability of psychological results and the value of the proposed methodological reforms. Replication rates obtained in newer large-scale replication projects (e.g., Klein et al. 2018) were consistent with the results of the aforementioned studies. Similarly pessimistic conclusions have been drawn from recent studies on the heterogeneity of psychological results. Heterogeneity is variance in the results of an experimental setup that exceeds the expected variance given the sampling error. Several recent studies examining multiple meta-analyses (see Linden 2019, Schauer and Hedges 2020 or Stanley, Carter, and Doucouliagos 2018) have found significant heterogeneity in the sampled experimental results. The heterogeneity was greater in studies that detected average or large effects and smaller in cases in which found effects were smaller. This heterogeneity remains largely unexplained and therefore constitutes another piece of evidence for the low reliability of psychological results. At the same time, inquiry into the psychological sources of the methodological problems has begun. Bishop (2020) argues that solutions to the crisis must consider deeper psychological causes of the crisis. According to Bishop, this includes confirmation bias, probabilistic fallacies, and asymmetric moral reasoning. All of those natural human tendencies translate into problematic scientific practices, therefore these influences should be accounted for. One of the proposed reforms is to encourage the use of triangulation, that is, using complementary approaches (e.g., experimental and observational studies) to study the same phenomena. Finally, it was recently shown that when the optimal methodology is implemented, high replicability is achievable. Protzko et al. (2020) presented the results of four replications on 16 new psychological results. The new results were obtained in methodologically optimal experiments. All experiments were preregistered,

well-powered, and methodologically transparent. Out of 64 replications, 55 were successful, which translates to 86% success rate, significantly higher than the disappointing results from other replication projects. This not only suggests that highly replicable and credible psychology is possible but also that the proposed reforms are effective.

In the rest of the paper, I will try to show that there are strong reasons to think that the reliability crisis of a similar nature and severity is present in forensic science.

# 3 Forensic Science

Forensic science aims at providing judicial evidence with a high degree of rigor and certainty. Forensic disciplines include DNA analysis, fingerprint examination, tool marks analysis, firearms analysis, and bite-mark analysis. The use of evidence provided by these disciplines is prevalent in courts around the world. Surprisingly, we know almost nothing about the error rates of these disciplines (see e.g., NRC 2009 or Koehler 2016a), and therefore, it is unclear how much confidence we should have in forensic results and verdicts based on them. In this section, I will describe in some detail Short Tandem Repeat Analysis, a type of DNA analysis, and ACE-V methodology, the leading methodology of fingerprints examination, and briefly discuss other disciplines.

## 3.1 DNA Short Tandem Repeat Analysis

DNA analysis is widely believed to be the most reliable of all forensic disciplines (see e.g., PCAST 2016, NRC 2009 or Butler 2009). Deoxyribonucleic acid (DNA) is an organic material present in most human cells, its chain-like structure contains the instruction of all essential functions of an organism such as growth or reproduction. Because of its function, DNA has a complex structure and varies between individuals. These features make a DNA trace suitable for identifying its source.

The most popular method of DNA analysis is the Short Tandem Repeats (STR) Analysis. STR is one of the most polymorphic, and therefore the most suitable for identifying fragments of DNA. STR analysis is more sensitive, takes less time, and provides more discriminating results than older methods (see e.g., Butler 2009). The process of analysis can be divided into six stages (for details see, Dash, Shrivastava, and Das 2020):

1. In the first stage, the organic traces containing DNA are collected. DNA can be obtained from almost all parts of the human body. The ideal tissue for this purpose is fresh blood, but other tissues such as skin, hair, semen, or urine can also be used. Thanks to amplification methods (see stage 4), even very old bones and teeth can be used as sources for DNA (see Reich 2018). The collection consists of extracting and separating part of the DNA-containing tissue (e.g., by cutting open the bone and collecting the marrow, which contains DNA) and placing it in a sealed container. Typically, one sample is collected at the crime scene and the other from the person of interest (e.g., a suspect).

2. DNA is extracted from the collected samples. The samples contain many components that are useless for identification. For example, out of all of the components of blood, only

leukocytes contain DNA, other blood components should be removed. The extraction in the case of blood proceeds as follows. The first step consists of the lysis of the cell membrane and the nucleus. This is typically done by adding a mixture of detergents and enzymes to the sample (e.g., sodium dodecyl sulfate and proteinase-K). The second step is to separate DNA from proteins, which is typically achieved by denaturation and the extraction of the DNA into an organic solution. DNA is not soluble in such a solution so it can be precipitated, washed, and dried. The isolated DNA can be then dissolved in water for long term storage. The procedure is similar in the case of other types of samples.

3. The next step is the quantification of purified DNA. To properly test a sample, we need to know how much DNA it contains. The amount of DNA in analogous samples may be different. For example, different samples of blood will contain different amounts of DNA depending on the amount of leukocytes in the sample. One technique used to quantify the amount of DNA is to use dyes that selectively binds only DNA particles. The bonded and activated particles of the dye emit a fluorescent light, the amount of which is proportional to the amount of DNA.

4. In this step, specific loci of the examined profile are amplified. Loci are fixed positions on a chromosome where particular genetic markers are located. Typically, the amplified fragment of DNA, which is later used for identification, is composed of 20 markers. Before the amplification, the amount of genetic material on those loci is too small to reliably detect the present alleles. The most widely used method of amplification is Polymerase Chain Reaction (PCR). It is an in vitro method of multiplying specific fragments of a given DNA. It consists of adding a primer designed to start the DNA synthesis of a specific region with the purified sample, and then heating it cyclically to inhibit the synthesis.

5. Detecting the presence of the alleles on the examined loci. This detection is typically done through the process of electrophoresis. Prepared samples are placed in the medium (e.g., gel or a narrow capillary) and moved by electric current. In the most recent version of the method–Capillary Electrophoresis, the amplified fragments of DNA travel through a sub-millimeter diameter capillary and are detected by a device that measures the time a given fragment takes to travel through the capillary and its light signature. Fluorescence intensity data are translated by dedicated software (e.g., GeneMapperID software) and presented as an electropherogram chart, where the detected alleles are presented as peaks on the graph.

6. The final step of the procedure is the analysis of the obtained results (see SWGDAM 2017 for details). First, the spikes on the charts obtained in the previous stage are interpreted by an expert on the basis of their height and position on the chart. A spike can be categorised as background noise, signs of impurities or alleles. During interpretation, experts rely on two thresholds: an analytical threshold and a stochastic threshold. The results are then compared with results from the analyses of a few control samples.

   After the interpretation is ready, both samples can be compared. There are three possible

decisions an expert can reach after comparing the samples. She can declare a match if the alleles are the same or the present differences can be explained away. She can declare that the samples do not match if there are any inexplicable differences, or declare that the result is inconclusive, for example, if a significant part of genotype is missing in one of the samples.

If the result is positive, then the expert needs to estimate the strength of the evidence provided by the match. This is done, typically, by estimating, on the basis of validated genetic databases and genetic principles, the probability that an unrelated person can be a source of the sample in question. The obtained probabilities are usually very low. For example, the probability of a random match reported during the investigation of Clinton–Lewinsky affair was on the order of 1 in 7.8 trillion (see Butler 2009, p. 70).

### 3.2 Fingerprints Analysis, the ACE-V methodology

Fingerprints are impressions left by friction ridge skin (the skin of the palms of the hands and fingers and the soles of the feet and toes) on touched surfaces. The structure of the friction ridge skin is believed to be persistent, highly discriminating and, until recently, unique (see e.g., Cole 2014). These features and the fact that fingerprints are recoverable from almost all surfaces make them useful for identification (see e.g., NIJ 2011, OIG 2011 or OSAC 2017).

Typically, fingerprints found at the crime scene are collected and compared (see e.g., Kasper 2015) with fingerprints collected from the suspect. ACE-V methodology is a leading methodology for fingerprints comparison (see SWGFAST 2002 or NIJ 2011). It consists of four stages corresponding to the letters in the acronym: analysis, comparison, evaluation, and validation. During the analysis stage, the examiner assesses the quality and features of the collected prints. The quality of a print is composed of factors such as completeness and clarity. If the examiner decides that the quality of a print is sufficient for a worthwhile further examination worthwhile, she proceeds to analyze its features. The features are divided into three levels of details.[1] The first level is the overall pattern of the ridges, there are three basic patterns: loops, whirls, and arches. The second level features are the configurations of individual ridges, which include where they start, how long they are, and how they are interconnected. The third level consists of the more fine-grained features of individual ridges such as edges, textures, and pore positions. Additional features such as scars or flexion creases are also taken into consideration on all three levels. After the analysis is completed and the quality of the fingerprint warrants further examination, the examiner proceeds to the comparison of both prints. During this stage, features from all three levels are compared. Indistinguishability among the features is not expected, because of possible distortions that could be caused by environmental factors. During the next step - the evaluation phase, the examiner makes a conclusion concerning the source of the fingerprints recovered from the crime scene. The three possible conclusions (see e.g., OSAC 2017) are: agreement (there is a sufficient level of agreement to determine that the source of both prints is the same), disagreement (there is a sufficient level of difference to determine that the sources of the two fingerprints are different), and inconclusive

[1] This distinction was introduced in (Ashbaugh 1999).

(there were not enough clear details to determine the source of the fingerprint). There is no set standard for similarity sufficient for identification. Instead, handbooks and guidelines tend to refer examiners to their experience:

> "Decisions are made throughout the perceptual process. A threshold, based on unique detail and expertise, is used to make decisions throughout the process." (NIJ 2011, p. 264)

### 3.3 Other Forensic Disciplines

Four other forensic disciplines are discussed in PCAST 2016: bitemark analysis, firearms analysis, footwear analysis, and microscopic hair analysis. In all these cases the procedure is somehow similar, the expert compares two samples, one collected from a crime scene (e.g., a footprint mark or a hair) and one collected from a person of interest (e.g., a shoe or heir sample). Typically, an expert first compares general features (such as type or color), and if those features are consistent, she proceeds to compare fine-grained features (such as wear marks or pigmentation of the cortex). If sufficient similarity is found among the features on both levels, the expert declares consistency. The methodologies employed in all those disciplines are analogous to the ACE-V fingerprints analysis. The main difference is that the reliability of those disciplines is rarely tested and if it is tested they typically score significantly worse than DNA or fingerprint analysis (see the next section). I will not discuss these methodologies in detail. Instead, I will focus on what we know about their reliability in the next section.

## 4 State of Forensic Science

It is notoriously difficult to assess the reliability of scientific results. Science is a complex enterprise conducted in various contexts and conditions, by individuals with various competencies and motivations. The same is true about forensic science. In general, there are at least three possible ways to test the reliability of results from a given scientific discipline:

**First Type** We can reexamine the existing scientific results. For example, we could try to replicate them (in the sense of large-scale replication projects à la Ebersole et al. 2016 or Collaboration 2015).

**Second Type** The reliability of scientific practice and by extension its results can be estimated by testing the performance of scientists in a setup similar to their everyday scientific practice.[2]

**Third Type** Finally, we can try to estimate (or rather argue for or against) the reliability of a given practice on the basis of theoretical considerations. This can take the form of presenting and defending

[2]This strategy of testing reliability is not common in meta-science. Scientific procedures are complicated and time-consuming and therefore it is hard to set up a suitable testing situation. One study utilizing a similar approach is (Dongen et al. 2019).

theoretical principles on which the reliability of a given procedure, and by extension its results, is based. A more by formal version of this approach is to model a given scientific practice to see how its known features influence the reliability of its results (such an approach was taken for example in Ioannidis 2005).

As we have seen, scientists are aware of the low degree of reliability in the psychological sciences, evident in large-scale replication studies. In these projects, scientists reexamined published results, so they belong to the first type of the reliability tests.

But what about forensic science? Sadly, despite the recommendation of NRC 2009, very little has been done to test the reliability of forensic science (see e.g., Koehler 2016a). This was also reasserted in PCAST 2016. In the next subsection, I will try to collect and sum up what little is known about the reliability of forensic science. Then, I will move toward a more speculative part of the paper in which I point out that all of the causes of the replicability crisis seem to be present in forensic science and therefore, our trust in forensic results should be limited.

## 4.1 What we know about the reliability of forensic science

Traditional justifications for the reliability of forensic science amount to the third type of reliability testing. These arguments appeal to the features of the procedures and the examined traces in order to argue that the results are likely to be reliable (see e.g., Gutiérrez-Redomero et al. 2010, Fagert and Morris 2015, Kücken and Champod 2013 or Rawson et al. 1984). There are also studies that fit the second type, most of them are dedicated to fingerprint analysis (see e.g., Pacheco, Cerchiai, and Stoiloff 2014 or Eldridge, De Donno, and Champod 2020). However, the studies of the first type are very rare in the forensic science literature (e.g., FBI 2015 or Dror and Hampikian 2011). In my discussion of the studies, I will rely on the PCAST 2016, a comprehensive report concerning the validity of the forensic disciplines conducted by the President's Council of Advisors on Science and Technology in 2016. In the report, the studies testing the reliability of forensic results have been collected and discussed. I will also add some recent studies which I was able to identify. I will start by discussing the studies of the third type.

DNA analysis is based on solid theoretical grounds. This convinced many that the results of DNA are reliable. According to the report, the DNA single-source and simple-mixture samples analysis are objective[3] and reliable. The validation studies demonstrated the accuracy and reliability of the procedures employed in DNA analysis (see e.g., Moretti et al. 2001 or Kimpton et al. 1996). The estimations of random match probability are based on sound, well-researched statistics describing the frequencies of specific alleles on investigated loci (see e.g., Shea, Niezgoda, and Chakraborty 2001). According to the report, additional subjective choices are involved in the interpretation of samples composed of DNA of three or more contributors. In such a case, each of the contributors may contribute one, two, or zero alleles on each locus, which makes the electropherogram hard to interpret, and therefore, the analysis is susceptible to confirmation bias as

[3]Objectivity is understood as following in the report: "By objective feature-comparison methods, we mean methods consisting of procedures that are each defined with enough standardized and quantifiable detail that they can be performed by either an automated system or human examiners exercising little or no judgment." (PCAST 2016, p. 47).

demonstrated in Dror and Hampikian (2011). Interestingly, this is one of the few studies which involve an attempt to replicate the results of a forensic procedure performed for the purpose of an actual court case (and therefore is a study of the first type). During the study, 17 independent experts were asked to reanalyze DNA evidence previously used in a court case. In contrast to the original examiner, experts participating in the study were not influenced by contextual information or biases, such as the details of the investigation, admission of guilt, etc. The results were clear, out of these 17 experts, just one reached a conclusion that is consistent with the original one. According to the authors, this suggests that the results of DNA mixture analysis depend on subjective choices of the examiner and therefore are susceptible to the biasing influence of contextual information.

The theoretical standing of fingerprint analysis is not as strong as that of DNA analysis. The procedure is intrinsically subjective, which means that the subjective judgments of the expert are involved in a substantial way. For example, experts have to decide if a given dissimilarity can be explained away by the possible distortions. This subjectivity is enforced by the flexibility of the ACE-V methodology. Neumann et al. (2014) compared the way fingerprint experts conduct their analyses. During the study, participants were asked to compare 15 pairs of fingerprint photos. In order to maximize the variability of obtained decisions, the authors used pairs of fingerprints, which were challenging cases due to the quality of the prints. Data concerning the observations and conclusions made by the participants were collected and used to assess which features of a given print, and to what degree have these features contributed to the decision of a particular expert. On this basis, the authors extrapolated how each expert understands the notion of similarity on which her decisions were based. The results show that examiners differ significantly in how they understand the similarity between a collected fingerprint sample and that of a suspect. Second, assumptions on which the discipline is based, such as the assumption that the fingerprints are unique or at least very discriminating (e.g., SWGFAST 2002), are not supported by existing evidence.

For these reasons, fingerprint analysis is classified in the report as a subjective discipline. The analogous problems are present in all other disciplines studied in the report, except single and double source DNA analysis. Therefore they are also classified as subjective disciplines. Moreover, the report states that the reliability of such subjective forensic procedures can only be established by means of type two reliability studies:

> "As discussed above, the foundational validity of a subjective method can only be established through multiple independent black-box studies appropriately designed to assess validity and reliability." (PCAST 2016, p. 91)

The report lists four such studies: Langenberg (2009), Tangen, Thompson, and McCarthy (2011), Ulery et al. (2011), and Pacheco, Cerchiai, and Stoiloff (2014). The last two are assessed to be methodologically most sound. The results of the experiments vary. (Langenberg 2009) and one of the experiments from (Tangen, Thompson, and McCarthy 2011) support the 0% false positive error rate. Error rates in (Ulery et al. 2011) were 0.17% and the results of the experiment from Pacheco, Cerchiai, and Stoiloff (2014) (when clerical errors are included) support 4.2% false positive error rate. False-positive error rates reported in a more recent study (Eldridge, De Donno, and Champod 2020) are consistent with the results of the older studies at 0.7%. The authors of the report

concluded that in light of the small number of studies, there is still a need for research, but the collected evidence suggests that fingerprint analysis is scientifically valid and can be used reliably.

Very little is known about the reliability of other forensic disciplines studied in the report. For example, in the case of bitemark analysis, the foundational assumptions such as the uniqueness of human dental profile, or that the human skin is a reliable medium for bitemarks are unsupported or contradicted by empirical results (e.g., Sheets, Bush, and Bush 2012 or Bush, Bush, and Sheets 2011). Similarly, type two reliability studies delivered disappointing results. For example, the results of Whittaker (1975) indicated that even in the best possible situation (the examined photographs of the marks were made immediately after the marks were made), experts made correct identifications in only 72% of the cases. The situation is similar in other subjective disciplines. Uniqueness assumptions are unsupported for hair analysis or footwear analysis and reliability studies are missing or deliver similarly high error rates. The reliability of hair analysis was tested in Houck and Budowle (2002). In the study, scientists attempted to replicate the results of hair analyses conducted by the FBI experts before the year 2000. They did so by conducting DNA analysis on the same hair samples. As reported in PCAST 2016, the false-positive rate, calculated on the basis of the results, was 11%.

## 4.2 The Causes for of the Crisis

In this section, I will point to some problems which decrease the reliability of forensic science and compare them to problematic issues present in academic science. I will start by describing problems that are specific to DNA analysis and subjective forensic disciplines. I will then discuss some of the problems which are common to all forensic disciplines.

### 4.2.1 The Problems of DNA analysis

DNA analysis is considered to be the most scientific and reliable out of all forensic disciplines. At the same time, there are at least three factors that may reduce the reliability of its results. First, there are still substantial degrees of flexibility involved in the method. Second, human errors can take place at any stage of the analysis. The probability of such errors is rarely tested and therefore hard to estimate. Finally, the increasing sensitivity of DNA analysis makes the interpretation of its results difficult.

Every stage of DNA analysis involves choices to be made by the scientist. There are many possible samples at the crime scene. There are a few ways to process the samples at each stage (see e.g., Dash, Shrivastava, and Das 2020) and there are at least two types of electrophoresis (capillary or gel electrophoresis). Finally, there is a lot of variability at the analysis stage. To see that, we can follow the latest version of Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (SWGDAM 2017) developed by Scientific Working Group on DNA Analysis Methods. The group consists of scientists representing federal, state, and local forensic DNA laboratories from the United States and Canada. According to the guidelines, many thresholds crucial for the analysis and interpretation of the results should be established internally

by the laboratory in which the analysis is conducted, for example:

> "The laboratory shall establish an analytical threshold. [...] The laboratory shall establish criteria to evaluate internal standards and allelic ladders. [...] The laboratory shall establish criteria for evaluation of controls. [...] The laboratory shall establish criteria to address the interpretation of non-allelic peaks." (SWGDAM 2017, p. 4)

The analytic threshold defines the height below which a peak is classified as an effect of background noise rather than a detected allele. It is easy to see how crucial this threshold is for the correct interpretation of the data. The other parameters from the above quote are equally crucial, and setting all of them to the right values is, according to the guidelines, the responsibility of the laboratory. This lack of standardized thresholds introduces flexibility to the procedure. As we have seen, there is substantial evidence suggesting that the flexibility of a scientific procedure decreases its reliability. Such flexibility introduces subjectivity to the analysis, and we know that in some cases experts can reach different conclusions on the basis of the same DNA sample. For example, Thompson (2005) discusses three experts who reached different conclusions on the basis of the same autoradiograph (another way of presenting DNA data). Moreover, we know that such degrees of freedom, introduced by increased methodological flexibility, are abused by experts in order to reach a positive result. For example, Thompson (2009) describes actual cases in which forensic experts abused the fact that there are no formal standards for inclusion. They did this in order to adjust the interpretation of borderline peaks such that the result would fit the profile of a suspect. Thompson points out that different interpretations of borderline peaks are not taken into consideration in calculating the probability of a random match. He adds that it's plausible that such practices occur in other forensic disciplines. A similar phenomenon was discussed in the context of psychology. Gelman and Loken (2019) shows that when methodological decisions are contingent on data, obtaining significant, but likely not replicable, results is very easy. In light of that the authors conclude that in such cases, a statistical significance of the obtained result is not a reliable measure of the results' reliability.

The second factor which may covertly decrease the reliability of DNA analysis results is human error. Human error can occur at any stage of an analysis and there are many well described cases of such errors (see e.g. Alexander 2015; Atkinson 2016; Iannelli 2016; Manna 2020; Shaer 2015). The sample may be contaminated (for a massive scale contamination see: Himmelreich 2009) or mistakenly classified as a single source sample while it is in fact a multi-donor sample (such example is discussed in Thompson, Taroni, and Aitken 2003). Similar errors can occur at other stages. These errors are typically unreported and their rates are not systematically studied:

> "Forensic practitioners in the U.S. do not typically report quality issues that arise in forensic DNA analysis. By contrast, error rates in medical DNA testing are commonly measured and reported." (PCAST 2016, p. 74)

The estimates of error rates in medical DNA analysis range between 0.61% and 0.31% (see: Kloosterman, Sjerps, and Quak 2014). As it was noted by Koehler (2016a), and by the PCAST (2016)

report, the probabilities of such errors are much higher than a typical random match probability (see also: Gill (2014) p. 56). This makes random match probability an unreliable indicator for the reliability of DNA results.

The third problem consists of the difficulties introduced by an increased sensitivity of DNA analysis. Since the introduction of PCR amplification, it is possible to analyze a minuscule amount of DNA–as small as a handful of cells. This possibility is used in analyzing trace DNA–very small amounts of DNA that has been left at the crime scene. In practice, trace DNA marks are often interpreted, for example, as left by the suspect touching some object at the crime scene or as particles of her shed skin. As described at length in Gill (2014), there are many ways in which trace DNA can be deposited at the crime scene which that does not incriminate the suspect.[4] Such scenarios include background contamination, the DNA of a suspect being deposited at the crime scene before the crime was committed, or contamination mediated by an investigator. The possibility of such contaminations makes the interpretation of trace DNA difficult, and as mentioned by Gill (2014), this has led to cases of wrongful arrests and convictions.

None of the above problems makes DNA analysis any less scientific. The lesson is rather that even scientific procedures are fallible and in some cases, as demonstrated by meta-scientists, they may consistently deliver unreliable results. Therefore, the scientific status of DNA analysis does not warrant its reliability. Moreover, we know that the random match probability commonly used as an estimate of the reliability of a DNA analysis does not represent other factors, which may be many times more detrimental to that reliability than the chance of a random match.

### 4.2.2 Problems of fingerprint analysis and other subjective forensic disciplines

All other disciplines discussed in PCAST (2016) (fingerprint analysis, bitemark analysis, firearms analysis, footwear analysis, and microscopic hair analysis) were found to involve a significant degree of subjective judgment. Procedures from all of those disciplines are flexible. They seem to rest on a series of subjective judgements by the scientist. These judgements concern whether there is enough similarity at each level of the comparison to declare that the examined objects are consistent. As we have seen in the case of fingerprint analysis, such flexibility was demonstrated in Neumann et al. (2014).

However, the reliability of the results produced by these procedures is often overstated. Until recently, fingerprint experts commonly claimed that they are able to uniquely identify the source of a given fingerprint, or that the error rate of ACE-V methodology equals zero (e.g., OIG 2011 p. 14). Similar claims, unsupported by scientific evidence, were made about other subjective forensic disciplines (see e.g., Garrett and Neufeld 2009).

Finally, we lack reliable statistics on how common are the features (e.g., patterns of the skin ridges) used by forensic experts while making identifications (see e.g., PCAST 2016 or NRC 2009). For this reason, it is unclear how to avoid overestimating the reliability of subjective forensic procedures.

[4] This problem was also discussed in Butler 2015 p. 458-9.

These problems undermine the reliability of forensic procedures. Moreover, we are often ignorant of how reliable those procedures are. Some of them may even be shown to be unreliable. For example, the examination of results from microscopic hair analysis conducted by FBI examiners before the year 2000 showed that at least 90% of those results and testimonies based on them contained errors. These errors included false identification and exaggerating the significance of the evidence (see FBI 2015).

### 4.2.3 Confirmation and Context Bias

Bias in forensic science is now a well-studied phenomenon. Surprisingly, the experimental studies of bias picked up relatively recently, after the publication of the NRC (2009). The report highlighted the possible detrimental effects of such biases and recommended further empirical studies. There is now an abundance of evidence showing that task-irrelevant information influence the decision of forensic scientists and by extension the results of forensic procedures (see e.g., Cooper and Meterko 2019). Such effects were demonstrated for the analysis of DNA mixtures (Dror and Hampikian 2011), fingerprints analysis (e.g., Smalarz et al. 2016), tooth mark analysis (Osborne et al. 2014), hair analysis (Miller 1987), and many other forensic disciplines (see e.g., Perry, Neltner, and Allen 2013, Nakhaeizadeh, Dror, and Morgan 2014, Kukucka and Kassin 2014).

Such experiments typically compare the verdicts of participants exposed to potentially biasing information with verdicts reached in absence of such information. For example, Smalarz et al. (2016) examines the effect of criminal stereotypes on the results of forensic procedures. Participants were asked to assess a pair of mismatching fingerprints. The fingerprints were supplemented with descriptions of the crime and the suspect. First, each participant received a randomly assigned description of a crime. The crime is either one associated with a demographic profile (child molestation associated with a Caucasian male) or not associated with any demographic group (identity theft). Second, participants received a randomly assigned description of the suspect that either fit the stereotypical child molester (Caucasian male), or does not fit (Asian female). After the examination of the descriptions and prints, participants were asked to conclude if they believe that the fingerprints match. The participants from the group exposed to the combination of descriptions fitting the stereotype were almost twice more likely to declared a false match (51.9% vs. 27.1%). The influence of task-irrelevant information on the forensic results is clearly detrimental to the reliability of forensic procedures. They should be as close to being responsive exclusively to the discriminating features of the investigated traces (shapes of ridges in fingerprints etc.) as possible, in order to provide a reliable and fair assessment of what a given object tells us about the investigated crime. This view is also reflected in the guidelines for the conduct of forensic disciplines (see e.g., NCFS 2016).

### 4.2.4 Funding bias

As mentioned in both NRC (2009) and PCAST (2016), forensic laboratories in the U.S. are typically funded by the prosecutor's offices, police departments, or the Federal Bureau of Investigation.

In some cases, experts are employed directly by these agencies (see e.g., FBI 2015). Both reports describe this as problematic. The performance of both prosecutor offices and police departments is assessed on the basis of their rates of successful convictions. This means that the funding institutions are interested in positive results that can be used to secure a conviction. The preference of the funding institution was shown to sway the results of both scientific experiments (see e.g., Wilholt 2008) and forensic procedures (see e.g., Murrie et al. 2013). There are, at least, three plausible mechanisms explaining the presence of such an effect in forensic science.

First of all, the prosecutors commission examinations, they therefore decide which laboratory and which experts would serve as analyst and witness in a given case. This makes it possible for the prosecutor to cherry-pick an expert who is more likely to deliver a positive decision. Such a preference may explain why some of the most notorious forensic experts are also some of the most active. For example, both Joyce Gilchrist and Fred Zain, two of the most notorious forensic experts, were popular among prosecutors to the point of being described as "prosecution superstars" (see Giannelli 2010). This effect promotes unreliable, biased experts, and therefore reduces the overall reliability of forensic results. Once again, an analogous and equally problematic phenomenon is present in academic science. The incentive structure of science favors poor research methodology that is more likely to deliver false-positive results (see Smaldino and McElreath 2016). Moreover, unreplicable and therefore unreliable results were shown to be more cited than replicable ones (see e.g., Serra-Garcia and Gneezy 2021).

The preference of funding bodies for positive results has also influenced the practice of experts. This influence can be both conscious and unconscious, and can influence the results in a few ways. The expert can make their methodological decisions with a positive result in mind. As we have seen, forensic procedures from all disciplines involved significant levels of flexibility, which can be exploited by a motivated expert to increase the probability of positive results. Similar covert practices, QRP are common in psychology and well studied in meta-science. There is substantial evidence suggesting that the presence of such practices decreases the reliability of experiments (see e.g., Simmons, Nelson, and Simonsohn 2011), and it seems that they potentially have the same effect in forensic science. I discussed an example of such a practice, a DNA expert adjusting the inclusion criteria to fit the known profile of a suspect in subsection 4.2.1 . Finally, the preference of the funder can influence the way forensic scientists present their results (see subsection 4.2.6).

### 4.2.5 Suppression of Exculpatory Evidence

A related problem stems from the fact that the prosecutor's office does not have to disclose to the defense which forensic examinations have been commissioned. Because of that, it is easy for them to conceal negative and exculpatory forensic results. Such misconducts are well-studied and there are many documented cases of it (see e.g., Jones 2010). Once again, there is an analogous problem in psychology called *publication bias* or the file drawer effect (see e.g., Scargle 2000). The publishing policy of the majority of scientific journals favors positive results (supporting the existence of the effect in question). Scientists are aware of this tendency and because of that, tend to be biased toward positive, publishable results. In effect, the published literature does not

represent the totality of collected evidence. In other words, it is easier to "prove" the existence of an effect in psychology than to disprove it. Given that the funding bodies for forensic procedures and prosecutors hold back exculpatory evidence, it seems plausible that the situation is similar in forensic science and forensic results are similarly skewed.[5] In effect, it may be easier of even significantly easier to incriminate the suspect than to exonerate him through forensic science.

#### 4.2.6 Cascades and Snowballs

Another factor that can further decrease the reliability of forensic science is the potential interactions between the afore mentioned detrimental effects. Dror et al. (2017) argued that the relations between different biases or biases at different stages of a forensic procedure can amplify the overall negative effect.

The effects of early-stage investigation (such as a collection of evidence or interrogation) are typically known to the experts who are involved in the later stages (like forensic analysis). In some cases, the same person is responsible for multiple tasks. Because of that, it is possible, and perhaps it should be even expected, that the information obtained at the early stages can bias the results of the later procedures, which can bias further activities, and so on. Dror et al. (2017) call this phenomenon *bias cascade*.

Secondly, the results (or preliminary results) of one forensic analysis can bias experts responsible for conducting other forensic examinations for the purpose of the same investigation. For example, the results of fingerprint analysis can be obtained by the expert responsible for the macroscopic hair analysis. This knowledge could motivate her to conduct her analysis in the way that is most likely to deliver a result consistent the decision of the fingerprints expert. This interaction is called the *bias snowball* (see also Edmond et al. 2015).

It is unknown if such effects actually occur, and as the authors acknowledged that needs to be established empirically.[6] On the other hand, the proposed mechanisms are plausible and can explain cases of wrongful conviction supported by faulty results of more than one forensic disciplines, or cases in which false forensic results track irrelevant factors such as confessions. Many such cases were documented by the Innocence Project (see IP 2020) and The National Registry of Exonerations (see NRE 2020).

#### 4.2.7 Misrepresentation

Even when a forensic procedure is concluded and the results are reached, there are still some ways in which the gathered evidence might be mishandled. For example, a testifying expert or the prosecutors can misrepresent the results while making their final statements during a hearing. This issue was investigated in (Garrett and Neufeld 2009). The authors analyzed 137 transcripts of trials

[5] It is unknown how often such misconduct takes place. Moreover, given how forensic research is organized, is not obvious if it can be ever reliably established.

[6] Similarly, it is not clear to what extent similar relations between different biases are present in academic science. As far as know, there is only anecdotal evidence present in the literature. For example, Stapel (2012) claims that some of his results based on fabricated data were later replicated by independent researchers.

which resulted in wrongful convictions (and later exoneration on the basis of new DNA evidence) and found that 60% of those trials involved invalid forensic testimony. In 80 out of 82 cases, the testimony presented the evidence as stronger than it in fact was. For example:

> "The auditors reached the unanimous conclusion that Officer LeBlanc realized at some point prior to trial that Cowans was excluded [as the source of fingerprints], but that he nevertheless concealed that fact in his trial testimony. Instead, Officer LeBlanc misrepresented to the jury that the latent print matched Cowans's. The auditors' conclusion was based on facts including: Cowans's exclusion was clear to every member of the review team; Officer LeBlanc had made correct associations and exclusions routinely in more difficult cases over the preceding four years; he made efforts to conceal other errors made in the same case; there were numerous inconsistencies in his testimony; and he intentionally used a method of showcasing the erroneous Cowans match evidence to the jury that not only made it more difficult for the jury to follow but was contrary to the preferred methods of fingerprint examiners and contrary to what Officer LeBlanc did with the other latent print in the same case." (Garrett and Neufeld 2009, p. 73-74)

The authors presented some evidence suggesting that such misrepresentations are equally common in trials that did not lead to wrongful convictions. Yarkoni (2020) claims that similar overselling of results is common in psychology and dramatically inflates false-positive rates.

### 4.3 How bad is it actually? Study of wrongful convictions in the US

In the third section, I discussed three ways of investigating how reliable the results of a given scientific procedure is. The fourth, less formal way would be to look for possible consequences of the hypothetical unreliability. Poor methodology in psychology results in the publication of highly implausible results (e.g., Bem 2011 or Bennett et al. 2010), or confused and contradictory literature on a given subject (e.g., Bishop 1990). What would be an analogous outcome in the case of unreliable forensic science? It would be wrongful convictions. The connection between unreliable procedures and wrongful convictions is not as direct as in the case of scientific results. Incorrect incriminating forensic results do not automatically result in wrongful convictions. On the other hand, it has been shown that the presence of incriminating results from forensic analysis significantly increases the probability of conviction (see e.g., Gould et al. 2013, Ling, Kaplan, and Berryessa 2021 or Smit, Morgan, and Lagnado 2018). Therefore it seems that we could gain some insight into the reliability of forensic science by investigating the rate and causes of wrongful convictions.

According to the United States National Registry of Exonerations (see NRE 2020) since 1989, 2674 prisoners have been exonerated. At least in 24% of the cases, false or misleading scientific evidence was present, and partly responsible for the wrongful conviction. This is extremely concerning. We know that in the last 30 years in the United States, at least 652 innocent people spent a significant time in prison at least partly because of faulty forensic results. Plausibly, the problem

is even worse, since it is highly likely that not all the wrongfully convicted were exonerated. The most recent study estimating the rate of wrongful convictions, Walsh et al. (2017) concluded that wrongful convictions constitute 11.6% of all convictions in the United States. According to Todd D. Minton and Zeng (2021), 2,086,600 persons were incarcerated in the US in 2019. Based on these two numbers we can estimate the number of wrongfully convicted prisoners in the US in 2019 as 242,045. This estimation clearly suggests how severe the problem of wrongful convictions is, and how much faulty forensic science contributes to such convictions.

# 5 How similar are both crises?

Phenomena analogous to all of the main causes of the replicability crisis are present in forensic science. All forensic disciplines exhibit substantial flexibility, there are substantial interests promoting positive results, forensic results are influenced by irrelevant factors and the results are often overstated. Additionally, PCAST (2016) claimed that all forensic disciplines, except single and double source DNA analysis, rely on subjective judgments of the responsible expert. We also know that one in four wrongful convictions (which ended up in exoneration) was supported by a faulty scientific result and there are estimations suggesting that wrongful convictions are disturbingly frequent. Interestingly, the structures of the replicability crisis and the crisis of forensic science are similar. As discussed in Andreoletti (2021), we can distinguish two groups of causes for the replicability crisis. First, features of the institutional structure of science promote unreliable results. A good examples of such a problem is the publication bias. Second, some of the scientific methods problematic or are used in a problematic way. An example of this is using underpowered experiments (see e.g., Vazire 2016). Accordingly, Andreoletti argues that to adequately address the replicability crisis, both types of problems need to be addressed. Similar dual-cause dynamics are present in forensic science. As we have seen, not only do all the methods used in forensic science involve aspects that are considered to be problematic (e.g., overstated reliability of the results or hidden methodological flexibility), but the institutional context in which the forensic results are produced and used are problematic (e.g., the influence of prosecutors on the forensic procedures or the suppression of exculpatory forensic evidence). In conclusion, all of the causes and symptoms of the reliability crisis are present in forensic science. Therefore, it is likely that forensic science is in a state of crisis with a magnitude similar to the replicability crisis in psychology. In other words, the reliability of forensic results is much lower than is commonly believed.

However, one might argue that some of the features of forensic science reduce the corrupting effects of the aforementioned methodological problems. One such factor would be the inherent difficulty of the involved tasks. In psychology (and other sciences) the most popular aim is to establish a general claim of correlation or causation (e.g., "cleanliness reduces the severity of moral judgments" from Johnson, Cheung, and Donnellan 2014). Such claims are established by means of experiments or observational studies and statistical inferences. At first glance, the aim of forensic procedures is different. A major goal of forensic procedures is to establish a singular claim, something like: "This specific fingerprint (one found at the crime scene) was produced by this

specific finger (namely one of the suspects)." Therefore it seems that the results of two enterprises are of different types. Moreover, the aim of psychology seems to be more difficult to achieve, generalizing on the basis of a limited number of observations is notoriously difficult and can be compromised in many ways. For example, the sample could be non-representative, or the statistical model could be ill-suited for testing the hypothesis in question. Therefore, it may seem that the goal of forensic science, positive identification, is easier to achieve. Consequently, one might think that the discussed problems have a lesser impact on forensic science and that its results could still have a sufficient degree of reliability.

At second glance, this defense of the reliability of forensic results falls short. The goal of forensic procedures is not only to identify the trace as originating from a suspect but also to exclude, with some degree of certainty, other people as the source of the trace. In light of that, it is clear that forensic scientists seek to support a stronger, more general claim (something like "This specific fingerprint (one found at the crime scene) was likely produced by this specific finger (namely one of the suspect), and it is unlikely that anybody else produced it."). Therefore, forensic procedures depend crucially on statistical results, such as describing the frequency of distinctive features of studied traces (e.g., skin ridges on the pointing finger forming a loop pattern). Because of that, they are indirectly susceptible to some of the same problems that face experimental science (e.g., unrepresentative sample). It is no longer clear that forensic aims are easier to achieve than scientific ones. Moreover, there are reasons to think that the situation in forensic science is even worse than in psychology. All subjective forensic disciplines involve a greater degree of flexibility than that of a typical methodology used in psychology. Additionally, publishing scientific articles typically involves peer review. A version of peer review (a second examiner internally replicating the analysis) is present in some forensic disciplines (such as DNA or fingerprint analysis), but not in others. Similarly, in some cases of forensic science, the required training for one to be considered an expert is alarmingly minimal. For example, microscopic hair experts have been certified after completing a two-weeks FBI-run course (see FBI 2015). Finally, even if the corrupting influence on the reliability of forensic results differs in some respects from those in psychology, it does not mean that they are any less problematic or that we cannot learn anything from the replicability crisis in psychology. The concerns over replicability have been raised in many other scientific disciplines such as biology (see e.g., Errington et al. 2021), physics (see e.g., Camerer et al. 2016), geography (see e.g., Kedron et al. 2021 or Sui and Kedron 2020), and a significant similarity can be found in both the causes and the proposed solutions.

The only piece of evidence still missing in the discussion of the methodological crisis in forensic science is a proverbial smoking gun. As we have seen in the second section, the results that convinced the majority of psychologists that their discipline is in crisis were large-scale replication studies. In such studies, by failing to replicate the results of published experiments, scientist showed that those experiments and by extension their conclusions are unreliable. This direct approach makes such studies not only reliable, but also persuasive in establishing the shortcomings of psychology. Unfortunately, there are almost no such (type one) studies conducted in forensic science. The usefulness of the most common-type two studies-which test the performance of an expert in experimental conditions may be limited in establishing the reliability of forensic results.

Most of the potential causes of unreliability are clearly not present in an experimental situation. Experts participating in such experiments (e.g., experiments from Pacheco, Cerchiai, and Stoiloff 2014) are not exposed to irrelevant information which could influence their decisions, and so confirmation bias is not a factor. Moreover participants typically know that they are participating in experiments, so they would want to maximise the reliability of their results. This sits in contrast to their everyday practice. Typically, experts are paid by the prosecutor's office and therefore incentivized to deliver incriminating results. Another problem with such studies, mentioned for example in PCAST 2016, is the possible presence of the Hawthorne effect (see e.g., Berthelot, Goff, and Maugars 2011). It is a well-documented effect that consists of the fact that observation tends to change the behavior of the participants in the study. All of the type two studies conducted up to date to test the reliability of the results of forensic science seem to be susceptible to those problems and therefore do not provide us with reliable estimates.

# 6 How should we proceed?

There are strong reasons to think that there is a crisis of reliability in forensic science. However, a crucial piece of evidence, the results of large-scale replication studies are still missing. In this section, I will discuss how to conduct such studies in forensic science and how to improve the reliability of forensic results. I will base my recommendations partly on proposals and results which emerged in the discussion surrounding the replicability crisis in psychology.

## 6.1 How to test for reliability of forensic science?

The most popular way to test the reliability of forensic results, type two studies, test the performance of forensic experts in unrealistic conditions, and therefore do not seem to be fully reliable. Contrary to PCAST (2016) (see page 9 for a quote), black-box type two studies are neither the only nor the best way to test the reliability of forensic procedures. To see that, let us return to the replicability crisis. The type of studies that were most effective in uncovering the replicability crisis in psychology were large-scale replication projects. Those type one studies examined a large number of published results, they present a clear, jarring, and bleak picture of psychology. Forensic science needs similar studies. Only by replicating actual forensic results, can we reliably assess their reliability. The theoretical studies will not be sufficient. The forensic experts' reports are notoriously brief and the presence of the reliability-reducing factors are typically omitted in them. For example, an expert will typically not report that her judgment was clouded by irrelevant information. Therefore, theoretical examination of the forensic reports and trial transcripts will likely result in overestimation on the reliability of those results. Such difficulties are not present in the replication approach. A replication study in the case of forensic science would involve experts reproducing **actual forensic results** by reanalyzing traces examined in the original analysis.[7] In

[7]It should be clear that I do not propose here to replicate the results of type two studies. Such an approach would suffer from difficulties described above.

designing such a study it is important to compose, as representative as possible, samples of actual forensic results. For example, the proportion of positive to negative results in a sample should resemble the analogous proportion in actual results. A second important question is how many results we should include. As far as I know, it is not yet clear how many results should a given sample include to be representative of a discipline in question (for some consideration of this problem see: Shaw et al. 2020). In light of that, perhaps 100 results can be treated as an acceptable minimal, in line with a number of psychological results included in Collaboration (2015). On the other hand, reproducing forensic analyses seems to be both faster and cheaper than reproducing psychological experiments. For example, during the black-box study reported in Eldridge, De Donno, and Champod (2020), 12,279 results of fingerprints analysis were collected. This suggests that replicating a similarly-sized sample of forensic results is feasible. In the study, we compare the sample of real forensic results subjected to all of the corrupting factors with the results of newly conducted analyses. Ideally, the second analysis should be conducted in an environment that is, as much as possible, free from such corrupting factors present during the original analysis (e.g., cognitive bias or funding bias). This is typically the case in psychological replications, which, for example, are published no matter what the result is, and therefore are not susceptible to publication bias (see Nosek and Lakens 2014). We can achieve similar results in forensic replications, for example, if they are funded by an independent institution which is unbiased toward specific results and does not provide experts with any irrelevant information. On the basis of the replication results, we can estimate the reliability of the original studies and the magnitude of the effects from corrupting factors present in the original study.

Type one replication study can be conducted in many ways. First of all, we can collect a number of analyses we want to replicate and assign each of them a single expert who would re-analyze the original traces. Such an approach is analogous to the one taken in Collaboration (2015). We can also ask more than one expert to re-examine each forensic result, as was done in Dror and Hampikian (2011). This would give us a more comprehensive picture by comparing the results of multiple analyses. Such a method can be extended to include multiple cases. We can compile a list of forensic problems composed, say, of two sets of fingerprints and present it to a number of experts. The analogous approach was taken in Klein et al. (2018).

What results should we expect from such a study? Let us consider the case of the fingerprints analysis. The estimations of reliability obtained in type two studies suggest that the replicability of the analysis is high. If the original procedure delivers the right result in $\approx 99\%$ of cases, then in $\approx 98\%$ of cases, two examiners should agree in their decision when they analyze the same prints. I suspect that the estimation of replicability we would obtain in type one studies will be significantly lower. We know that forensic science suffers from many problems which contribute to the replicability crisis in psychology, and it is likely that type two studies overestimate the reliability of fingerprints analysis.

Additionally, we may want to re-analyze forensic results from a particular forensic discipline with methodologies from another discipline. As discussed in section 4.1, a replication of this type was conducted by the FBI to test the reliability of microscopic hair analysis. Its results were re-assessed by conducting DNA analysis on the same hair samples in Houck and Budowle (2002). The

study showed a large discrepancy in the results between the two methods. In 11% of cases, results of their analysis were not supported by the results of the DNA analysis. These results suggest that microscopic hair analysis is not as reliable as it was previously believed. A similar approach can be used to re-assessed the reliability of other forensic disciplines.

Plausibly, running a large-scale replication project in forensics may encounter many bureaucratic and technical difficulties. The samples may be nonexistent or classified, and judges and prosecutors may be inclined to block such attempts in order to defend convictions. Discussing these problems goes beyond the scope of the paper and the competence of the author, but the lesson from psychology seems to be that it may be the most efficient way to obtain trustworthy estimates of the reliability of forensic results.

## 6.2 How to improve the reliability of forensic science?

We have all the reasons to believe that the reliability of forensic science is overestimated. In light of that, it makes sense to think about possible reforms that would improve reliability. Both the replicability crisis in psychology and the reliability crisis in forensic science are complex phenomena, it is plausible that there is no easy way to solve them. Many proposals were made in an attempt to improve the reliability in both disciplines and some analogies can be found. At the same time, I will argue that although some of the ideas from the replicability crisis literature have not yet been discussed in forensic science, they could be equally useful in this new context.

### 6.2.1 Transparency and Openness

As we have seen, one of the main elements of the metascientific response to the replication crisis was highlighting the importance of transparency in the scientific process. We trust scientific results because they are justified by reliable scientific procedures. But how can we trust a given result if we do not know the methodological details of the study which supported it? This is why transparency is now seen as one of the defining features of the scientific process. This crucial role of transparency can be traced back to the ethos of modern science including the norm of organized skepticism defended in Merton (1973). But how is transparency is understood in science? A comprehensive and up to date conceptualization is provided in a recent consensus-based checklist presented in Aczel et al. (2019). The checklist consists of a list of questions asking if each of the main features of a given experiment (like statistical analysis, sample size, or measures of interest) were preregistered, whether they are explicitly described and motivated in the paper, and whether the data software etc. were made publicly available. If all of those conditions are satisfied, then the given scientific procedure is transparent and we can judge for ourselves if the delivered result is trustworthy. Transparency does not by itself make the scientific procedure or its results any more reliable, but it is necessary for assessing their reliability.

Is transparency seen as equally crucial in forensic science? Surprisingly, the answer seems to be "no". The transparency of procedures employed by forensic scientists is rarely mentioned in both NRC (2009) and PCAST (2016). For example, in PCAST (2016), transparency is narrowly

understood as disclosing the expected error rates of the employed method. Such error rates are not typically disclosed by laboratories in the US, and in line with the recommendation from Kloosterman, Sjerps, and Quak (2014), authors of (PCAST 2016) recommend that forensic laboratories should disclose them. It is easy to see that following this recommendation falls short of achieving transparency in the sense developed in (Aczel et al. 2019). Similarly, the reports of forensic procedures in which the results are presented are typically short and sketchy. For example, the guidelines by the Friction Ridge Subcommittee from the US-based Organization of Scientific Area Committees for Forensic Science (see OSAC 2017) are purely negative. The experts are warned against using misleading formulations such as "exclusion of all others", "100% certainty", or "zero error rate" but the guidelines do not outline which details of the analysis must be explicitly mentioned in the report. Consequently, an acceptable report of fingerprint analysis might be very brief and therefore the fingerprints analysis is, in many cases, far from being transparent. Recently, the issue of transparency of forensic analysis gained some attention in the forensic literature (see e.g., Chin, Ribeiro, and Rairden 2019; Kruse 2013; Passalacqua, Pilloud, and Belcher 2019). Kruse (2013) has convincingly argued for two claims: First, that the Bayesian approach to quantifying the uncertainty of forensic results, widely used by forensic experts in Sweden, contributes to the transparency of the forensic process. Second, forensic results of different disciplines can be easily combined in this framework, which promotes the intersubjectivity of such results. Interestingly, the discussed American guidelines directly dismiss this approach:

> "Reported conclusions shall be expressed as the opinion of the examiner. The examiner has a level of personal confidence associated with the accuracy and reliability of this conclusion; however, this personal level of confidence cannot be objectively measured. For this reason, certainty shall not be reported in absolute terms and should not be reported numerically." (see OSAC 2017 p. 8)

Adequate discussion of this controversy goes beyond the scope of the paper. On the other hand, given the arguments presented in Kruse (2013), the Swedish approach seems to be superior.

Chin, Ribeiro, and Rairden (2019) and Chin, McFadden, and Edmond (2020) argue persuasively that forensic science will benefit from adopting reforms developed in the open science framework (see e.g., National Academies of Sciences and Medicine 2018) to improve the transparency of science. According to the authors, both forensic methodological research as well as forensic procedures should be conducted in an open and transparent way. All aspects of forensic procedures should be explicitly described and preregistered, and all the collected data should be made public. The authors mentioned three advantages of such an approach: it would increase the reliability of forensic science, make the use of forensic science in court fairer, and accelerate the transition from subjective forensic methods to more objective ones. Implementing those recommendations would undoubtedly go a long way in improving the transparency of forensic science.

### 6.2.2 Inclusive reliability assessments

As we have seen, the reliability of forensic procedures is commonly overestimated. Even in the case of the most reliable forensic discipline, DNA analysis, the most commonly used estimate of error rates-–random match probability is too optimistic. As we have seen, laboratory errors and the misinterpretation of results are far more salient reasons for false or misleading results from DNA analyses. In light of that, using the probability of random match as a proxy for the probability of error in the DNA analysis seems to be, in the best case scenario, a misleading idealization. The situation is similar in cases of other forensic disciplines, the error probabilities are either missing or based on highly idealized estimates (e.g., obtained in black box studies in the case of fingerprint analysis).

Consequently, forensic science needs new estimates for the reliability of forensic procedures. In order to provide a fair representation of that reliability, the estimates need to take into account the effects of all methodological problems identified in the literature.

### 6.2.3 Reducing the Flexibility

The methodological flexibility of a given procedure has been determined to be one of the main factors in reducing the reliability of the procedure. This seems to be similar in forensic science. Disciplines that exhibit greater degrees of flexibility are considered to be less reliable. Consequently, the authors of NRC (2009) and PCAST (2016) have recommended developing currently used subjective and therefore flexible methods into more rigid, objective ones. The ongoing developments in forensic methodology will plausibly reduce flexibility and improve reliability (see e.g., Neumann, Evett, and Skerrett 2012 or Neumann et al. 2014).

There are many ways to reduce methodological flexibility. As already mentioned, preregistration have became popular in psychology (see e.g., Nosek and Lakens 2014). Preregistration consists of describing the methodology (data collection method, employed moderators, analysis method, etc.) of the experiment before it is conducted. This prevents the scientist from changing those features mid-experiment in order to artificially increase the chance of getting positive results. As we have seen, preregistration has also been proposed for forensic science, and it would plausibly amend some of the diagnosed problems. For example, if an inclusion threshold is predefined before the start of a DNA analysis, it is no longer possible to adjust the threshold in light of the obtained data, which is problematic (see Thompson 2009).

Second, the thresholds crucial for a given procedure should be fixed in order to restrict the flexibility and therefore possible misuses. An example of fixing a threshold in response to the replication crisis is statistical power (see Vazire 2016). Perhaps the thresholds crucial for forensic disciplines such as analytical, stochastic, or inclusion thresholds in DNA analysis or the required degree of similarity in fingerprints analysis can be fixed in a similar way. This would reduce the flexibility of those procedures and the possibility of methodological misuse (e.g., QRP), and therefore improve the reliability of forensic procedures. At the same time, it should be noted that reducing flexibility may be costly and should be done carefully. Not all of the possible values for

the thresholds in question will be equally successful. For example, if the value of the inclusion threshold is too low, the number of false-negative results will be high. In light of that, the values that are selected need to be validated empirically and fine-tuned.

### 6.2.4 Restricting the effects of the external influences

The most popular response to the issue of confirmation bias in forensic science is *Sequential Unmasking* (Krane et al. 2008 or Robertson and Kesselheim 2016). It consists of hiding information that are inessential at a given stage of the analysis from the expert. If an expert does not know, for example, whether a suspect confessed to a given crime, or what the result of some other analysis was, this knowledge cannot bias her decisions. Plausibly, such a strategy will be effective in mitigating the confirmation bias, but not the funding bias. It would be difficult to hide from an expert who his employer is and where her interests lie.

### 6.2.5 Changing the incentive structure

The most obvious solution to the problem of funding bias and related issues is to change the incentive structure of a given discipline. In the case of academic science, we can create an independent institution that would be responsible for organizing experiments commissioned by interested parties (see e.g., Krimsky 2006). Such an institution would be responsible for choosing the scientific laboratory for the study and assessment of whether the conducted experiments are methodologically acceptable. Given that the institution does not care if the results are positive or negative, it is able to manage experiments in a way conducive to them being reliable and objective. If properly structured and run, such an institution would be capable of counter the funding bias. The need for institutional oversight in forensic science has been acknowledged in the literature (see e.g., NRC 2009), and steps toward introducing such an institutional oversight in the United States have been made. The main aim of the National Commission on Forensic Science, active from 2013 to 2017, was to improve the reliability of forensic science. The role of the commission was advisory: for example, it produced documents describing the conditions for proper forensic practice (see e.g., NCFS 2016). Sadly, the funding of the commission was discontinued (see Costakes 2017) and it is unclear if attempts to establish institutional oversight for forensic science in the US will be continued.

## 7 Conclusion

Before concluding, I will discuss a few other ideas. Some of them are present in meta-science, which can be helpful for increasing the reliability of forensic science, but may be unrealistic or dificult to implement.

## 7.1 Unreliable until proven reliable

One of the fundamental principles of common law is the presumption of innocence. The defendant in a criminal case is considered to be innocent until the prosecution is able to prove beyond reasonable doubt that she has committed the crime in question. Surprisingly, in the case of the incriminating forensic results, it seems to be the other way around. The expert typically presents their results as scientifically valid and reliable, and it is the responsibility of the defense to undermine this reliability in front of a jury. As we have seen, experts are unaware of many of the problems mentioned above and are typically convinced that their results are reliable (see e.g., Murrie et al. 2019) or even perfectly reliable (see e.g., Koehler 2016a). Moreover, they are likely to oversell their results, and both defense lawyers and judges are typically inefficient at opposing questionable forensic results (see Garrett and Neufeld 2009). In light of that, perhaps incriminating forensic evidence should be assessed in line with of the presumption of innocence: the evidence of guilt is not reliable until it is proven to be reliable. Such a solution may be paralyzing for the use of scientific evidence in court. As we have seen, even in the case of the most reliable forensic discipline: DNA analysis, there are still factors such as human error or the misuse of methodological flexibility, which reduce the reliability of a forensic result to an unknown degree. In light of that, demonstrating the reliability of some forensic results could be virtually impossible and therefore, too much to ask. On the other hand, some forensic disciplines may be unreliable, so such reliability cannot be demonstrated and so results of those disciplines should not be used. In any case, in light of the common failures of forensic science, it seems to be justified to place the burden of demonstrating the reliability of a given result on the forensic expert. Additionally, such a skeptical approach to scientific evidence is clearly in line with the spirit of the presumption of innocence.

## 7.2 Consistency of results

There seems to be a consensus in meta-science that we can learn very little from the results of a single experiment. In order to establish a stable and reliable effect, we need to conduct multiple studies (see e.g., Ioannidis 2005). If the results of a significant majority of those experiments converge, then we can be reasonably confident that a supported hypothesis is true. Once a sequential unmasking is implemented and forensic scientists work independently, a similar approach may be effective in forensic science. When the procedures are independent, we gain additional evidence by combining positive results from different analyses. In light of that, we can validate the results of a forensic procedure by replicating it. As in the case of the replicability study described in section 6.1, a result can be replicated by different experts conducting the same type of analysis. Alternatively, one can try to validate the results of one forensic analysis by conducting the analysis from another discipline on the same sample. For example, results of fingerprint analysis can be validated by analyzing trace DNA contained in the samples. Reversely, a lack of such convergence between results should be treated as a sign that something went wrong, so perhaps positive results should not be trusted. To some degree, such cross-validation has been implemented in forensic science. For example, ACE-V method involves a validation step. On the other hand, results which oppose the

initial identification made by forensic experts who are employed by the prosecutor's office are typically disregarded. For example, during the trials which lead to the wrongful convictions of Lana Canen (see Chinn 2012) and Richard Jackson (see Possley 2019), exonerating results of fingerprint analysis inconsistent with initial identifications were disregarded by the jury. If the replication-like approach was implemented more consistently, and if the disagreeing results were to be taken into consideration, perhaps the wrongful convictions could have been avoided.

### 7.3 Lack of evidence

Given that the sensitivity of the DNA analysis has increased to an impressive degree, perhaps we can start to treat the lack of DNA evidence as evidence that a suspect was not present. Such an approach would go against the "absence of evidence is not evidence of absence" principle (see e.g., Gill 2014), but given that DNA profile can be now reconstructed from minuscule amounts of DNA, in some cases, it may be improbable that a person could commit a crime without leaving a behind DNA trace. If such probability can be estimated, then the absence of evidence can have some evidential value. The arguments from the absence of evidence have been reconstructed in the Bayesian framework, and it was shown that some instances of such arguments are valid (see e.g., Oaksford and Hahn 2004). The predictions of the Bayesian model are supported by the results of experiments testing the intuitions of participants (see e.g., Hahn and Oaksford 2005 or Hsu et al. 2017). Such an approach has been also used in science. For example, Karl Popper famously claimed that each unsuccessful attempt to falsify a theory corroborates it (see e.g., Thornton 2018). Similarly, an unsuccessful replication of a given hypothesis makes it, in the eyes of scientists, less plausible. At the same time, we should be careful in using such evidence, since there are possible ways to prevent the depositing of DNA-containing particles or ways to remove them from the crime scene. Those possibilities have to be taken into consideration in assessing the epistemic import of an absence of evidence. On the other hand, those possibilities do not make the negative evidence evidentiary inert and the "absence of evidence is not evidence of absence" principle should be reconsidered in light of the developments in scientific methodology and the sensitivity of forensic methods (see also Thompson 2018 and Taroni et al. 2019).

### 7.4 Self-correction

Finally, the replication crisis is sometimes interpreted as a failure of scientific self-correction. Self-correction is considered to be one of the crucial features of scientific practice. Scientists make mistakes, but sooner or later those mistakes should be detected, exposed, and repaired. There is a lot of self-correcting mechanisms in place in science, such as peer-review or replication projects. On the other hand, it is not clear to what degree they are successful (see e.g., Romero and Sprenger 2019). Such mechanisms, with the exception of validation steps present in some disciplines, are largely missing in forensic science. The forensic result seems to be set in stone. Defense lawyers may try to undermine old results, but this appeal mechanism seems to be ineffective. Discussing the ways in which self-correction can be implemented in forensic science goes beyond the scope

of the paper, but I will try to point to one way such an attempt can be integrated into the ongoing struggle against wrongful conviction. Organizations such as the Innocence Project work daily to exonerate those who have been wrongfully convicted. One way to undermine wrongful convictions is to undermine the evidence on which they were based. Therefore, it might be both possible and useful to establish a publicly funded "scientific wing" for such an organization. It would selectively re-analyze forensic results used in court. We know that forensic evidence has a great impact on verdicts and that there is a multitude of question-raising convictions that are potentially wrongful. Such an approach would not provide us with a representative estimate of replicability, the scientist would replicate "suspicious results" which likely does not constitute a representative sample. On the other hand, such a strategy could be effective in targeting misleading evidence and supporting exonerations.

In my article, I argued that the reliability of forensic results is overestimated. I showed that all of the problems which caused the replicability crisis in psychology are present in forensic science. In light of this, it is at highly plausible that forensic science is in a state of methodological crisis analogous to the replicability crisis. I also discussed steps that need to be taken in order to amend the situation. First, to obtain a reliable estimate of the reliability of forensic science we need to conduct large-scale replication studies. Second, I discussed some of the proposals for increasing the reliability of forensic results such as implementing sequential blinding or changing the incentive structure of forensic science.

**Funding** There are no conflicts of interest. This research has been funded by the Polish National Science Centre [grant number 2016/22/E/HS1/00304].

**Acknowledgements** I would like to thank Rafał Urbaniak, Mattia Androleti, Jan Sprenger, Gustavo Cevolani, Davide Coraci, Weronika Majek, Patryk Dziurosz-Serafinowicz, Pavel Janda, Paweł Pawłowski, Robert Różański and the anonymous referees for their useful comments.

## References

Aczel, Balazs, et al. 2019. "A consensus-based transparency checklist." *Nature Human Behaviour* 4 (1): 4–6.

Alexander, Keith L. 2015. "Prosecutors criticize D.C. crime lab's handling of some DNA evidence." https://www.washingtonpost.com/local/crime/dc-prosecutors-criticize-city-crime-labs-handling-of-some-dna-cases/2015/03/05/b5244f88-bea4-11e4-b274-e5209a3bc9a9_story.html.

Andreoletti, Mattia. 2021. "Replicability Crisis and Scientific Reforms: Overlooked Issues and Unmet Challenges." *International Studies in the Philosophy of Science* 33 (3): 135–151.

Ashbaugh, D.R. 1999. *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology*. Practical Aspects of Criminal and Forensic Investigations. Taylor & Francis.

Atkinson, Khorri. 2016. "Austin Scrambles with Fallout of Closed DNA Lab." `https://www.texastribune.org/2016/07/30/more-questions-austin-police-department-lab/`.

Baker, M. 2016. "1,500 scientists lift the lid on reproducibility." *Nature* 533:452–454.

Bakker, Marjan, A. van Dijk, and J. Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7:543–554.

Bem, Daryl J. 2011. "Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect." *Journal of personality and social psychology* 100 (3): 407.

Bennett, C. M., et al. 2010. "of Serendipitous and Unexpected Results Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon : An Argument For Proper Multiple Comparisons Correction."

Berthelot, J., B. Le Goff, and Y. Maugars. 2011. "The Hawthorne effect: stronger than the placebo effect?" *Joint, bone, spine : revue du rhumatisme* 78 4:335–6.

Bishop, D. V. M. 1990. "How to increase your chances of obtaining a significant association between handedness and disorder." PMID: 1701771, *Journal of Clinical and Experimental Neuropsychology* 12 (5): 812–816.

Bishop, Dorothy. 2019. "Rein in the four horsemen of irreproducibility." *Nature* 568:435–435.

Bishop, Dorothy VM. 2020. "The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture." PMID: 31724919, *Quarterly Journal of Experimental Psychology* 73 (1): 1–19.

Bush, M., P. Bush, and H. Sheets. 2011. "Statistical Evidence for the Similarity of the Human Dentition." *Journal of Forensic Sciences* 56.

Butler, John. 2009. *Fundamentals of Forensic DNA Typing*. Academic Press.

Butler, John M. 2015. *Advanced Topics in Forensic DNA Typing: Interpretation*. Elsevier Academic Press: San Diego.

Camerer, Colin, et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351:1433–1436.

Camerer, Colin, et al. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2:637–644.

Chin, J. 2014. "Psychological Science's Replicability Crisis and What It Means for Science in the Courtroom." *Psychology, Public Policy and Law* 20:225–238.

Chin, Jason M., Rory McFadden, and Gary Edmond. 2020. "Forensic science needs registered reports." *Forensic Science International: Synergy* 2:41–45.

Chin, Jason M., Gianni Ribeiro, and Alicia Rairden. 2019. "Open forensic science." *Journal of Law and the Biosciences*: 255–288.

Chinn, Jeff. 2012. "Fingerprint Expert's Mistake Leads to Wrongful Conviction in Indiana." Accessed: 2022-02-11. https://californiainnocenceproject.org/2012/10/fingerprint-experts-mistake-leads-to-wrongful-conviction-in-indiana/.

Cole, Simon A. 2014. "Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States." *Law, Probability and Risk* 13, no. 2 (): 117–150.

Collaboration, Open Science. 2015. "Estimating the reproducibility of psychological science." *Science* 349 (6251).

Cooper, G., and Vanessa Meterko. 2019. "Cognitive bias research in forensic science: A systematic review." *Forensic science international* 297:35–46.

Costakes, Ariana. 2017. "Department of Justice to End National Commission on Forensic Science." https://www.innocenceproject.org/department-justice-ends-national-commission-forensic-science/.

Dash, Hirak, Pankaj Shrivastava, and Surajit Das. 2020. *Principles and Practices of DNA Analysis: A Laboratory Manual for Forensic DNA Typing*.

Department of Justice (U.S.), Oversight and Review Division. 2011. *A Review of the FBI's Progress in Responding to the Recommendations in the Office of the Inspector General Report on the Fingerprint Misidentification in the Brandon Mayfield Case*. U.S. Department of Justice, Office of the Inspector General, Oversight / Review Division.

Dongen, Noah van, et al. 2019. "Multiple Perspectives on Inference for Two Simple Statistical Scenarios." *The American Statistician* 73 (): 328–339.

Dror, I., and Greg Hampikian. 2011. "Subjectivity and bias in forensic DNA mixture interpretation." *Science & justice : journal of the Forensic Science Society* 51 4:204–8.

Dror, Itiel, et al. 2017. "The bias snowball and the bias cascade effects: Two distinct biases that may impact forensic decision making." *Journal of Forensic Sciences* 62 ().

Ebersole, Charles R., et al. 2016. "Many Labs 3: Evaluating participant pool quality across the academic semester via replication." Special Issue: Confirmatory, *Journal of Experimental Social Psychology* 67:68–82.

Edmond, Gary, et al. 2015. "Contextual bias and cross-contamination in the forensic sciences: the corrosive implications for investigations, plea bargains, trials and appeals." *Law, Probability and Risk* 14:1–25.

Eldridge, Heidi, Marco De Donno, and Christophe Champod. 2020. "Testing the Accuracy and Reliability of Palmar Friction Ridge Comparisons – A Black Box Study." *Forensic Science International* (): 110457.

Errington, Timothy M., et al. 2021. "Investigating the replicability of preclinical cancer biology." *eLife* 10.

Fagert, Michael, and Keith Morris. 2015. "Quantifying the limits of fingerprint variability." *Forensic Science International* 254:87–99.

Federal Bureau of Investigation. 2015. "FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90 Percent of Cases in Ongoing Review." https://www.fbi.gov/news/pressrel/press-releases/fbi-testimony-on-microscopic-hair-analysis-contained-errors-in-at-least-90-percent-of-cases-in-ongoing-review.

Garrett, Brandon, and Peter Neufeld. 2009. "Invalid Forensic Science Testimony and Wrongful Convictions." *Virginia Law Rev.* 95 ().

Gelman, Andrew, and E. Loken. 2019. "The garden of forking paths : Why multiple comparisons can be a problem , even when there is no " fishing expedition " or " p-hacking " and the research hypothesis was posited ahead of time."

Giannelli, Paul. 2010. "Scientific Fraud." *Criminal Law Bulletin*.

Gigerenzer, Gerd. 2004. "Mindless statistics." *The Journal of Socio-Economics* 33 (5): 587–606.

Gill, Peter. 2014. *Misleading DNA Evidence: Reasons for Miscarriages of Justice*. Elsevier Science.

Gould, J., et al. 2013. "Predicting Erroneous Convictions: A Social Science Approach to Miscarriages of Justice." *Criminology eJournal*.

Gutiérrez-Redomero, Esperanza, et al. 2010. "Distribution of the minutiae in the fingerprints of a sample of the Spanish population." *Forensic science international* 208 (): 79–90.

Hahn, Ulrike, and Mike Oaksford. 2005. "How convinced should we be by negative evidence." *Proceedings of the 27th Annual Conference of the Cognitive Science Society* ().

Heide, Rianne de, and Peter Grünwald. 2017. "Why optional stopping is a problem for Bayesians."

Himmelreich, Claudia. 2009. "Germany's Phantom Serial Killer: A DNA Blunder." http://content.time.com/time/world/article/0,8599,1888126,00.html.

Houck, M., and B. Budowle. 2002. "Correlation of microscopic and mitochondrial DNA hair comparisons." *Journal of forensic sciences* 47 5:964–7.

Hsu, Anne S., et al. 2017. "When Absence of Evidence Is Evidence of Absence: Rational Inferences From Absent Data." *Cognitive Science* 41 (S5): 1155–1167.

Iannelli, Jerry. 2016. "BSO Crime Lab Could Be Mishandling Crucial DNA Evidence, Whistleblower Says." https://www.browardpalmbeach.com/news/bso-crime-lab-could-be-mishandling-crucial-dna-evidence-whistleblower-says-7881208.

*Innocence Project (IP) website*. 2020. https://www.innocenceproject.org/all-cases/. Accessed: 2020-10-21.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2, no. 8 (): 55–69.

Johnson, David, Felix Cheung, and M. Donnellan. 2014. "Does Cleanliness Influence Moral Judgments? A Direct Replication of Schnall, Benton, and Harvey (2008)." *Social Psychology* 45 (): 209.

Jones, C. 2010. "A Reason to Doubt: The Suppression of Evidence and the Inference of Innocence." *Journal of Criminal Law & Criminology* 100:415–474.

Kasper, S.P. 2015. *Latent Print Processing Guide*. Elsevier Science.

Kedron, Peter, et al. 2021. "Reproducibility and replicability: opportunities and challenges for geospatial research." *International Journal of Geographical Information Science* 35:427–445.

Kimpton, C., et al. 1996. "Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification." *ELECTROPHORESIS* 17.

Klein, Richard A., et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–490.

Kloosterman, A., M. Sjerps, and Astrid Quak. 2014. "Error rates in forensic DNA analysis: definition, numbers, impact and communication." *Forensic science international. Genetics* 12:77–85.

Koehler, J. 2016a. "Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences."

— . 2016b. "Intuitive Error Rate Estimates for the Forensic Sciences."

Krane, Dan E., et al. 2008. "Sequential Unmasking: A Means of Minimizing Observer Effects in Forensic DNA Interpretation." *Journal of Forensic Sciences* 53.

Krimsky, S. 2006. "Science in the private interest: has the lure of profits corrupted biomedical research?" *IEEE Technology and Society Magazine* 25:10–11.

Kruse, Corinna. 2013. "The Bayesian approach to forensic evidence: Evaluating, communicating, and distributing responsibility." *Social Studies of Science* 43 (5): 657–680.

Kücken, Michael, and Christophe Champod. 2013. "Merkel cells and the individuality of friction ridge skin." *Journal of Theoretical Biology* 317:229–237.

Kukucka, Jeff, and S. Kassin. 2014. "Do confessions taint perceptions of handwriting evidence? An empirical test of the forensic confirmation bias." *Law and human behavior* 38 3:256–70.

Langenberg, G. 2009. "A Performance study of the ACE-V process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process." *Journal of Forensic Identification* 59 (): 219–257.

Linden, Audrey Helen. 2019. “Heterogeneity of research results: New perspectives on psychological science.”

Ling, Shichun, Jacob Kaplan, and Colleen M. Berryessa. 2021. “The importance of forensic evidence for decisions on criminal guilt.” *Science & Justice* 61 (2): 142–149.

Manna, Nichole. 2020. “A scientist in Fort Worth’s crime lab says rules were broken. Now a judge wants answers.” `https://www.star-telegram.com/news/local/fort-worth/article245756430.html`.

Merton, Robert King. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Miller, Larry S. 1987. “Procedural Bias in Forensic Science Examinations of Human Hair.” *Law and Human Behavior*.

Moretti, Tamyra, et al. 2001. “Validation of Short Tandem Repeats (STRs) for Forensic Usage: Performance Testing of Fluorescent Multiplex STR Systems and Analysis of Authentic and Simulated Forensic Samples.” *Journal of forensic sciences* 46 (): 647–60.

Murrie, D., et al. 2019. “Perceptions and estimates of error rates in forensic science: A survey of forensic analysts.” *Forensic science international* 302:109887.

Murrie, Daniel, et al. 2013. “Are Forensic Experts Biased by the Side That Retained Them?” *Psychological science* 24 ().

Nakhaeizadeh, Sherry, I. Dror, and R. Morgan. 2014. “Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias.” *Science & justice : journal of the Forensic Science Society* 54 3:208–14.

National Academies of Sciences, Engineering, and Medicine. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: The National Academies Press.

— . 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press.

National Commission on Forensic Science. 2016. *Ensuring That Forensic Analysis is Based Upon Task-Relevant Information*.

National Institute of Justice (U.S.) 2011. *The Fingerprint Sourcebook*.

National Research Council. 2009. *Strengthening forensic science in the United States: A path forward*. 1–328.

Neumann, C., I. Evett, and James E. Skerrett. 2012. “Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm.” *Journal of The Royal Statistical Society Series A-statistics in Society* 175:371–415.

Neumann, C., et al. 2014. “Improving the Understanding and the Reliability of the Concept of ”Sufficiency” in Friction Ridge Examination.”

Nosek, Brian, and Daniël Lakens. 2014. "Registered Reports A Method to Increase the Credibility of Published Results." *Social Psychology* 45:137.

Oaksford, Mike, and Ulrike Hahn. 2004. "A Bayesian Approach to the Argument From Ignorance." *Canadian journal of experimental psychology = Revue canadienne de psychologie expérimentale* 58 (): 75–85.

Organization of Scientific Area Committees for Forensic Science. 2017. *Guideline for the Articulation of the Decision-Making Process Leading to an Expert Opinion of Source Identification in Friction Ridge Examinations*.

Osborne, Niki, et al. 2014. "Does contextual information bias bitemark comparisons?" *Science & Justice* 54 ().

Pacheco, Igor, Brian Cerchiai, and Stephanie Stoiloff. 2014. "Miami-Dade research study for the reliability of the ACE-V process: Accuracy & precision in latent fingerprint examinations" ().

Pashler, Harold, and Eric–Jan Wagenmakers. 2012. "Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?" *Perspectives on Psychological Science* 7:528–530.

Passalacqua, Nicholas V., Marin A. Pilloud, and William R. Belcher. 2019. "Scientific integrity in the forensic sciences: Consumerism, conflicts of interest, and transparency." *Science & Justice* 59 (5): 573–579.

Peels, Rik. 2019. "Replicability and replication in the humanities." *Research Integrity and Peer Review* 4 (1): 2.

Perry, B., Matthew Neltner, and Timothy S. Allen. 2013. "A Paradox of Bias: Racial Differences in Forensic Psychiatric Diagnosis and Determinations of Criminal Responsibility." *Race and Social Problems* 5:239–249.

Possley, Maurice. 2019. "RICHARD JACKSON." Accessed: 2022-02-11. `https://www.law.umich.edu/special/exoneration/Pages/casedetail.aspx?caseid=3318`.

President's Council of Advisors on Science and Technology. 2016. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*.

Protzko, John, et al. 2020. *High Replicability of Newly-Discovered Social-behavioral Findings is Achievable*.

Rawson, R. D., et al. 1984. "Statistical evidence for the individuality of the human dentition." *Journal of forensic sciences* 29 1:245–53.

Reich, David. 2018. *Who We Are and How We Got Here. Ancient DNA and the New Science of the Human Past*. Oxford University Press.

Robertson, C. T., and A. Kesselheim. 2016. "Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science, and Law."

Romero, Felipe. 2018. "Who Should Do Replication Labor?" *Advances in Methods and Practices in Psychological Science* 1 (4): 516–537.

Romero, Felipe, and Jan Sprenger. 2019. "Scientific Self-Correction: The Bayesian Way."

Scargle, Jeffrey. 2000. "Publication bias: the "File-Drawer" problem in scientific inference." *Journal of Scientific Exploration* 14:91–106.

Schauer, Jacob M., and Larry V Hedges. 2020. "Assessing heterogeneity and power in replications of psychological experiments." *Psychological bulletin*.

Scientific Working Group on DNA Analysis Methods. 2017. *Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories*.

Scientific Working Group on Friction Ridge Analysis. 2002. *Friction Ridge Examination Methodology for Latent Print Examiners*.

Serra-Garcia, Marta, and Uri Gneezy. 2021. "Nonreplicable publications are cited more than replicable ones." *Science Advances* 7 (21): eabd1705.

Shaer, Matthew. 2015. "The False Promise of DNA Testing." `https://www.theatlantic.com/magazine/archive/2016/06/a-reasonable-doubt/480747/`.

Shaw, Mairead, et al. 2020. "Measurement practices in large-scale replications: Insights from Many Labs 2." *Canadian Psychology/Psychologie canadienne* 61 ().

Shea, Brendan, Stephen Niezgoda, and Ranajit Chakraborty. 2001. "CODIS STR loci data from 41 sample populations." *Journal of forensic sciences* 46 (): 453–89.

Sheets, H., P. Bush, and M. Bush. 2012. "Bitemarks: distortion and covariation of the maxillary and mandibular dentition as impressed in human skin." *Forensic science international* 223 1-3:202–7.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological science* 22 (11): 1359–1366.

Smalarz, Laura, et al. 2016. "The perfect match: Do criminal stereotypes bias forensic evidence analysis?" *Law and human behavior* 40 4:420–9.

Smaldino, Paul E., and Richard McElreath. 2016. "The natural selection of bad science." *Royal Society Open Science* 3 (9): 160384.

Smit, N., R. Morgan, and D. Lagnado. 2018. "A systematic analysis of misleading evidence in unsafe rulings in England and Wales." *Science & justice : journal of the Forensic Science Society* 58 2:128–137.

Stanley, T. D., Evan C. Carter, and Hristos Doucouliagos. 2018. "What Meta-Analyses Reveal About the Replicability of Psychological Research." *Psychological Bulletin* 144:1325–1346.

Stapel, Diederik. 2012. *Ontsporing*. Prometheus Amsterdam.

Sui, Daniel Z., and Peter Kedron. 2020. "Reproducibility and Replicability in the Context of the Contested Identities of Geography." *Annals of the American Association of Geographers* 111:1275–1283.

Swazey, J., M. Anderson, and K. Lewis. 1993. "Ethical Problems in Academic Research." *American Scientist* 81:542–553.

Tangen, J., Matthew B. Thompson, and Duncan J. McCarthy. 2011. "Identifying Fingerprint Expertise." *Psychological Science* 22:995–997.

Taroni, F, et al. 2019. "More on the question 'When does absence of evidence constitute evidence of absence?' How Bayesian confirmation theory can logically support the answer." *Forensic science international* 301 (): e59–e63.

*The National Registry of Exonerations (NRE) website*. 2020. `https://www.law.umich.edu/special/exoneration/Pages/about.aspx`. Accessed: 2020-10-21.

Thompson, Nicholas, William C.; Scurich. 2018. "When does absence of evidence constitute evidence of absence?" *Forensic Science International vol. 291* 291 ().

Thompson, W. 2005. "Subjective interpretation, laboratory error and the value of forensic DNA evidence: Three case studies." *Genetica* 96:153–168.

— . 2009. "Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation." *Law, Probability and Risk* 8:257–276.

Thompson, William C, Franco Taroni, and Colin G G Aitken. 2003. "How the probability of a false positive affects the value of DNA evidence." *Journal of Forensic Sciences* 48 (1): 47–54.

Thornton, Stephen. 2018. "Karl Popper." `https://plato.stanford.edu/entries/popper/`.

Todd D. Minton, Lauren G. Beatty, and Zhen Zeng. 2021. "Correctional Populations in the United States, 2019 – Statistical Tables." `https://bjs.ojp.gov/library/publications/correctional-populations-united-states-2019-statistical-tables`.

Ulery, Bradford, et al. 2011. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences of the United States of America* 108 (): 7733–8.

Vandenbroucke, J. 1988. "Passive smoking and lung cancer: a publication bias?" *British Medical Journal (Clinical research ed.)* 296:391–392.

Vazire, Simine. 2016. "Editorial." *Social Psychological and Personality Science* 7 (1): 3–7.

Walsh, Kelly, et al. 2017. *Estimating the Prevalence of Wrongful Convictions*. `https://www.ncjrs.gov/pdffiles1/nij/grants/251115.pdf`. Accessed: 2020-10-21.

Whittaker, D. 1975. "Some laboratory studies on the accuracy of bite mark comparison." *International Dental Journal* 25:166–71.

Wicherts, J., et al. 2016. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking." *Frontiers in Psychology* 7.

Wilholt, Torsten. 2008. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science Part A* 40 (1): 92–101.

Witte, Erich, and Frank Zenker. 2017. "From Discovery to Justification: Outline of an Ideal Research Program in Empirical Psychology." *Frontiers in Psychology* 8.

Yarkoni, Tal. 2020. "The generalizability crisis." *Behavioral and Brain Sciences* (): 1–37.