A Submission to the Open call for evidence:

Use of evidence generated by software in criminal proceedings

Alistair Kelman[1] and James Christie[2]

Executive Summary

To address the issues of **Admissibility** and **Reliability** of evidence from computers and automated systems we have a proposal which could separate the technical issues of admissibility from the day to day operation of advocacy in the courtroom.

Our proposal would embed Bayesian reasoning into an updated version of the **Seven Statement Test**, (proposed by Alistair Kelman in his book "*The Computer in Court*" in 1982) thereby establishing a method of statistical inference that uses prior knowledge and new evidence to update probabilities of hypotheses.

The Bayesian reasoning task would be performed by auditors and technical staff (and not by lawyers who are not needed for this kind of work). Any organisation which wished to present computer evidence would be required to have a Bayesian Analysis of the reliability of its evidence (based upon the elements of the updated Seven Statement Test) which would be prepared from audit and security records and would give rise to a probability which is a number between Zero and One where Zero is considered 0% and One is considered 100%.

In this paper we briefly explain Bayesian Analysis using an example from the Post Office Horizon case to show the importance of "***edge***" cases, consider the operation of safety-critical systems, set out a possible Statutory provision and illustrate how the relevant provision could be implemented at minimal cost to the public purse through a change in the law in respect of

[1] https://www.linkedin.com/in/alistairkelman/

[2] https://www.linkedin.com/in/clarotesting/

Directors and Officers Policies of insurance which we believe would be supported by the insurance industry.

Why would Bayesian Analysis help us to assess computer evidence?

Bayes Theorem is a powerful means of helping people make sense of complex and confusing statistical problems. We instinctively understand Bayes at a simple level, but non-statisticians can struggle when it comes to understanding and applying it in messy and complex real world situations.

To illustrate the operation of Bayes we present an example from the key civil case in the Post Office scandal, the Horizon Issues trial[3]. Dr Robert Worden, the Post Office expert witness, had argued that the odds were hugely against the postmasters who had brought the case against the Post Office having genuinely suffered from system errors. His argument was based on the overall reliability of Horizon for the Post Office's purposes. The system worked satisfactorily for millions of transactions a day, therefore the odds against one of the 555 postmasters in the litigation having suffered a problem were about **five million to one.** As Justice Fraser said in his judgment dismissing Worden's statistical argument *"it is a little more sophisticated than that, but not by very much"*[4].

Patrick Green KC neatly exposed the fallacy of Worden's argument. He asked the witness to work his way through the calculations that a random person he met was called Penny Black. Worden made appropriate and sensible assumptions about it being a woman's name, the prevalence of the surname Black and the first name Penny, and arrived at an estimate of 500,000 to 1. Green then asked Worden what the odds would be of one person in a group of 50 being called Penny Black and the odds of two bearing that name. 10,000 to one, and 100 million to one were the answers.

In 2015 the Royal Mail celebrated the 175th anniversary of the first adhesive postage stamp, which became known as the Penny Black, and invited everyone with that name to attend a dinner. Green asked Worden to revisit his calculation of the odds of meeting a Penny Black if he had also attended the event.

Worden objected that the Penny Blacks had been invited and were not randomly selected individuals. If you knew about the special event then "specific knowledge overrides probability theory, when you have that specific knowledge", as Worden put it.

Worden obtusely could not see how this was relevant to the 555 postmasters who had brought the litigation against the Post Office. The point was, however, clear to Justice Fraser. The 555 were not randomly selected. They were not a representative sample of the whole branch estate. They had come forward because they claimed to have suffered problems.

[3] We are grateful to Patrick Green KC, who made a very clear and persuasive argument when he questioned# the Post Office's expert witness Dr Robert Worden about his use of statistics see Transcript of the Horizon Issues trial, day 18, June 11 2019. https://www.postofficetrial.com/2019/06/horizon-trial-day-18-transcript.html

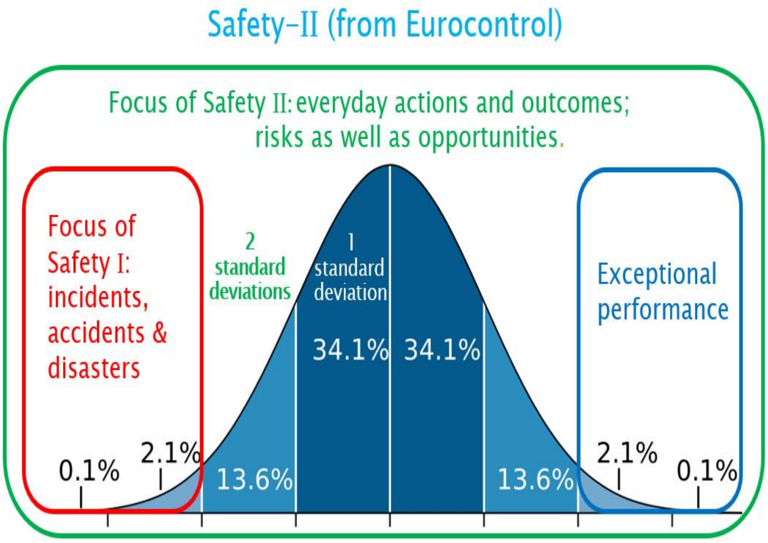
[4] Bates v Post Office Ltd (No 6: Horizon Issues) Rev 1 [2019] EWHC 3408 (QB). Paragraph 826. https://www.judiciary.uk/wp-content/uploads/2019/12/bates-v-post-office-judgment.pdf

Horizon might well have been working adequately as a whole for the purposes of the Post Office's corporate accounts and the corporation's management of the branches, but that did not mean it worked perfectly at every branch all of the time. Reasonable observers would recognise that as a logical fallacy even if they did not know about the many problems with the system. It is possible, and reasonable, to assume that the system was both generally reliable and also capable of causing deficits at individual branches.

When massive, complex systems like Horizon are working as designed they are unlikely to cause controversy. If controversy does arise, leading to civil litigation or criminal prosecution, it is far more likely to involve the extremes of system behaviour. It is therefore important to think about the possibility that a system was operating at its extremes at some locations, with some users, under unusual conditions, with the system suffering unpredictable minor, localised failures.

Safety Critical Systems

It is useful to consider how experts working with safety critical systems look at this issue. Traditionally, following the Safety-I model, investigators and engineers learned how to prevent accidents and disasters by examining and learning from these incidents after they happened. Routine performance received less attention. Indeed it was often ignored.


Safety–II (from Eurocontrol)
Focus of Safety II: everyday actions and outcomes; risks as well as opportunities.
Focus of Safety I: incidents, accidents & disasters
2 standard deviations
1 standard deviation
Exceptional performance
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%

The limitation of this approach was that the very success of engineers working with safety critical systems inevitably reduced the material from which they could keep learning. The Safety-II approach[5] entails examining and learning from what usually goes right. What do users do to keep systems safe? How is the system managed so that it never strays into the extremes where accidents happen?

When legal professionals work with computer evidence they should remember that they are not necessarily looking at random data from a system working normally. It is distinctly possible that they are looking at data and events that would be in the red box of the Safety-II diagram. It is unsafe to assume that the normal probabilities applying to everyday activities in the centre of that curve (a normal distribution) are relevant to the case in hand. It would be safer to assume greater probabilities of problems and failure. Glib assertions of evidence reliability, as offered by the Post Office and its expert witness, should not be accepted without supporting secondary evidence.

[5] Eurocontrol (European Organisation for the Safety of Air Navigation). 'From Safety-I to Safety-II: A White Paper', 2013. https://skybrary.aero/sites/default/files/bookshelf/2437.pdf

What might be considered a reasonable estimate of 99% reliability in normal operation could fall towards 50%, or even lower, at the extremes in the absence of supporting secondary evidence to justify the assertion that the system was operating in the safe zone.

This secondary evidence would be the equivalent of the "specific knowledge" that Patrick Green and Robert Worden were talking about in the Horizon Issues case. It would not be reasonable, or practical, to assume that all computer evidence is suspect simply because the case is contested and the party that did not provide the evidence is arguing that it must be flawed. Only if there is no supporting secondary evidence would that be a reasonable stance. But what form should supporting evidence take so that courts can place greater confidence in the computer evidence brought forward? That requires reassurance for courts that the evidence has not come from the extreme end of that Eurocontrol diagram and has been generated by systems that were being managed responsibly in the safe territory of the middle.

Alistair Kelman's Seven Statements[6] proposal provides an answer, and it can be combined with a statistical Bayesian analysis to give confidence that computer evidence can safely be considered reliable.

Bayesian Analysis - in greater depth

Bayes' Theorem is a highly effective, and widely accepted, means of combining two sets of probabilities to give a valuable answer that would not be available intuitively or if one looked at only one set.

A simple way to explain Bayes is to use the example of testing for diseases. If the test for a disease X produces a correct diagnosis 90% of the time when a patient has the disease, and a false diagnosis 30% of the time if the patient is healthy, what is the probability that a patient with a positive result has the disease? It is not 90%. You have to know more, i.e. you have to know the prevalence in the general population. If 20% are suffering from X and 80% are healthy then you would expect the following results from testing 100 people.

True positive	**= 18 (i.e. 20% times 90%)**
False positive	**= 24 (i.e. 80% times 30%)**
True negative	**= 56 (i.e. 80% times 70%)**
False negative	**= 2 (i.e. 20% times 10%)**

18 people have X but 42 tested positive (18+24). Therefore the probability that someone who tested positive actually has the disease is 18 divided by 42, approximately 43%.

The probability that we could work on if we knew only the test's accuracy was qualified by the additional specific knowledge that the base rate in the whole population was much lower, and

[6] Alistair Kelman. 'Revolutionizing Court Evidence: The Bayesian Approach'. 2025. https://revolutionizing-court-ev-cakwhl3.gamma.site

the test would therefore generate more false positives than true positives. In Bayes terminology the test accuracy is called the prior probability, and the more valuable prediction of whether a particular positive test is correct is called the posterior probability.

This example is easy to understand, but Bayes Theorem comes into its own with much more difficult and obscure problems where obvious answers are not available and a simple reliance on common sense will not work.

There are many benefits to using Bayes Theorem. It has been criticised for relying on subjective estimates of probability. That is true to a certain extent, but its handling of assumptions are a strength rather than a disadvantage.[7]

Conventional treatment of reliability statistics in a court case would entail consideration of a single figure, the estimate of reliability of the evidence in question. In isolation that can be misleading. Estimates might be based on ill-informed guesswork by witnesses lacking expertise. Probability figures might be relevant elsewhere and applied out of context.

The probability assumptions in a Bayes Theorem can be tailored to permit a more rounded and helpful picture of what is known about the evidence. This can, and should be done, by people with the appropriate technical and statistical expertise. Once made explicit the implications of these assumptions and estimates can be explored and challenged.

In particular, Bayesian analysis recognises that there are four possible outcomes in any test of evidence, whereas a conventional non-Bayesian approach focuses on two outcomes.

A Bayesian approach lets us quantify the expected rates for true positives, false positives, true negatives, and false negatives that are consistent with our assumptions. A simpler traditional approach that relies on binary positive/negative results, or the overall accuracy rate, doesn’t explicitly reveal all four outcomes. Such an approach might emphasise true positives and false negatives if we’re trying to detect a condition, or true negatives and false positives if we are trying to rule it out. It therefore lacks the clarity of Bayesian analysis.

Tests are invariably better in one direction than the other. In the example above the accuracy rate was 90% if the subject did have the disease, but only 70% if they were clear.

That would be the case with tests to provide an estimate of the reliability of computer evidence. It would be harder to detect reasons to confirm the evidence is reliable than it would be to detect reasons to dismiss it as unreliable. If we want to know whether evidence should be considered reliable we want to understand how many false positives and false negatives we are dealing with. That requires a Bayesian approach.

Judges and tribunals when considering admissibility of evidence always need to consider the reliability of that evidence in a two-fold manner. It would cause serious problems for the administration of justice if false computer evidence was stamped out by such unrealistically stringent tests that reliable evidence was being rejected wholesale. Consequently in every case the two-fold approach which leads to four outcomes is:

- How many unreliable pieces of evidence will be admitted? *true positive, false positive*

[7] For a detailed analysis of Bayes see ***"The Theory that would not die"*** by Sharon Bertsch McGrayne - ttps://www.amazon.co.uk/Theory-That-Would-Not-Die/dp/0300188226

- How many reliable pieces of evidence will be rejected? *true negative, false negative*

In order to understand all four outcomes (*true positive, false positive, true negative, and false negative*) we have to use Bayesian analysis so that we are not merely guessing on the basis of potentially harmful assumptions that we are not making explicit.

The Seven Statement Test when combined with Bayesian analysis enables judges and tribunals to understand the implications of admitting or rejecting evidence on a more considered basis than is currently possible. It would offer the additional specific knowledge about the probabilities of system reliability that courts need. It would give the providers of computer evidence a powerful incentive not only to build and manage systems responsibly but to ensure they can demonstrate how they have done so. This would allow them to show that they are keeping their systems out of the extremes shown in the Safety-II diagram and in the territory where it can be safely assumed that the probabilities point towards reliability of evidence. The alternative for these providers of evidence would be the galling prospect of seeing their evidence repeatedly rejected because they are unable to give the courts any confidence that it is accurate.

A Statutory provision to implement Bayesian Analysis in computer records

We propose the following as a means of addressing this issue.by a Statutory Amendment to the evidence legislation which is as follows::

> **In any proceedings, a statement in a document produced by a computer shall not be admissible as evidence of any fact stated therein unless it is shown –**
>
> 1. **that a Bayesian Analysis of all the risk factors that could lead to the statement being inaccurate are below 0.XX;**
> 2. **In cases where the Bayesian Analysis is below 0.XX then the judge has a discretion to admit the statement as evidence after a *voir dire hearing on the admissibility of the evidence.***
> 3. **In cases where the Bayesian Analysis is below 0.XX then any D&O Policy protecting Directors and Officers of the company deposing the computer evidence shall be Void;**

Our recommendation is that Parliament should set the acceptability level of probability for reliability of evidence for use in criminal cases. It would require a figure of say XX% probability i.e 0.XX

If the Bayesian Analysis gave a level of probability for reliability of evidence which was lower than 0.XX but the authorities wished to proceed with a trial then there would be a preliminary hearing (referred to as a *voir dire* hearing) where technical evidence was presented and cross-examined to establish how reliable the evidence was before the actual trial commenced.

In Alistair Kelman's book he said:

A document containing the Seven Statements would be extremely lengthy but much of it could and, we believe, should, be prepared by an organisation using computers prior to any incident requiring the organisation to go to law or to assist in a prosecution. The first six of the Seven Statements could be kept in draft form on file. It would then be a simple and inexpensive process to finally add the Seventh Statement and engross the document attaching any relevant computer printouts to it as exhibits.

The Bayesian Analysis would be prepared by technical consultants who would calculate it from a weighting of the updated Seven Statements.

Below we set out in a table the Original Seven Statements and a possible framework for updating them with a note on the Bayesian Analysis

The Seven Statements		
Original Seven Statements	**Updated Seven Statements**	**Bayesian Analysis**
Statement One should deal with the qualifications and experience of the person in charge of the computer system. This is to establish that he is capable of swearing such a document.	Today this should be a Board Member with explicit responsibility for ensuring that the company/department maintained accurate records. If the company/department outsourced its records maintenance to a cloud service or a service provider the Service Provider would provide this statement. Here I recommend that consideration is given to the FCA City Code and the US Sarbanes-Oxley Act (SOX) which protects American investors and helped rebuild trust in the financial markets after Enron and other major accounting scandals. The US approach is more prescriptive than the FCA one but facts suggest that this is an area where SOX would give rise to better protection	Factor in risks associated with any lack of seniority and experience in the deponent. In many automated systems this may be immaterial. However if there is any subjective aspect which could undermine the validity of the deponent then it would be included here.

Statement Two should consist of a description of the computer system with reference to each of the components in the system by brand and model number, e.g. a Kamikaze DDB7 with the Asthma 2.6 operating system running custom written payroll programs.	Here the Board Member or Service Provider would outline the key components that make up the computer system by brand, model number and configuration. This could be in a Wiki format whereby there were links to each element which made up the system.	Factor in risks associated with any brand, model number, components and configuration. If for example there were known problems with floating point operations with particular components that led to latent errors this would be included here. These might be associated with temperature sensitivity - something which might have to be addressed in a later statement.
Statement Three, a long statement, should deal with the quality of the individual components by reference to the development time involved in their creation. For example reference could be made here to any technical literature or manuals which were used, giving the number of man hours involved in their original development. Manufacturers of quality products would gladly assist in producing technical evidence of this kind.	A modern Statement Three would be a fully hypertext reference document which would identify the exact sources of information. Modern Artificial Intelligence (AI) systems could assist lawyers and judges in interrogating these references to discover latent errors and flaws.	A scoring system would produce the probability of latent errors and flaws being in the system

Statement Four should deal with the testing and documentation standards applied to any custom written software. If the software had been bought-in, the software house, if reputable, should be willing to provide information on its testing and documentation standards.	This Statement would consider the testing and documentation standards used in the writing of any custom written software and would be specified in accordance with a recognised documenting architectural tool. Thus the Statement regarding the documentation standards applied to any custom written software might be written in **DITA** (Darwin Information Typing Architecture) which is an XML-based standard that enables the creation of modular, reusable, and adaptable content. Or the documentation standards applied to any custom written software might be **DocBook** which is another XML-based standard that provides a schema and a set of tools for creating technical documentation. Similarly this Statement might be written in **Markdown** which is a lightweight markup language that allows developers to write plain text documents that can be converted into HTML or other formats. All of these documentation architectural tools have their own advantages and can be used for different types of documentation, such as books, articles, reference manuals, tutorials, README files, wikis, blogs, and notes.	The software testing criteria would generate a probability of reliability of the evidence
Statement Five should deal with the procedures for logging updates to the software and the qualifications of the subordinate staff involved in the computer system.	Here a modern **Statement Five** should deal with the way in which software is updated and maintained. This today would deal with overnight security updates and patches plus hashing and end to end encryption. Modern Artificial Intelligence (AI) systems could assist lawyers and judges in interrogating these references to discover latent errors and flaws.	The software updating procedures would generate a probability of reliability of the evidence

Statement Six should deal with the physical and electronic security features of the installation.	A modern **Statement Six** would also consider how computers could be remotely accessed both legitimately (by support staff working with an operator) and illegitimately.	The physical and electronic security features of the installation would generate a probability of reliability of the evidence
Statement Seven should indicate how the particular computer printout came into existence and what it purports to show. In this section the person in charge can say that no faults manifested themselves during the material time which would indicate to him that the computer evidence could not be relied upon.	There would be no material change in **Statement Seven** which would be sworn at the time the evidence was required.	Here specific additional factors (e.g. temperature sensitivity) could be addressed and explained away in giving a further probability component

Pre-written affidavit or deposition - a standard report

The basis of the Seven Statement test turns upon the fact that anyone wishing to rely on computer evidence in court proceedings submits a prescriptive affidavit or deposition in seven parts. This affidavit or deposition provides a window on the reliability of the submitted computer evidence because it gives detailed information on all aspects of the computer evidence in a standard format that would enable lawyers and judges (and juries) to form reasonable conclusions on the reliability of the evidence being presented to the court. It would enable non-technical lawyers to raise questions and gain answers so that we could be sure that only reliable computer evidence was considered by the court. Our conclusion is that it should be a regulatory requirement for a major company or a government department to comply with a modern Seven Statement test. Six of the Seven Statements regarding the reliability of the computer evidence could be pre-written and held as detailed draft internal audit reports in a standardised format as recommended by auditors and engineers. There would be an audit duty to maintain these drafts as being accurate and up-to-date as part of the normal bookkeeping requirements of all businesses and organisations. Consequently, it would only require the preparation of the Seventh of the Seven Statements, the detailed evidence which was being submitted and the associated report regarding its production, that would need to be drafted and filed with each deposition of computer or internet evidence.

The Third clause in the Statutory provision - D&O policies

As an incentive to ensure compliance with these provisions we recommend that there is a third provision in the evidence legislation. This would be that if any company or organisation failed to have a comprehensive and up-to-date Seven Statement Test statement available for use in reliance upon its computer records, the automatic consequence of this would be that any Directors and Officers Liability insurance cover taken out by the company or by its directors and officers would be void. This rule would apply to private companies as well as public companies. There is precedent for measures of this kind: In 1966 the Fire, Auto and Marine Insurance Company (FAM) collapsed leaving an estimated 400,000 motorists without insurance coverage. This led to a requirement under an amendment to the law regulating insurance companies that if an insurer was undercapitalized and fails to comply with financial regulations regarding having sufficient assets to meet claims against it, the policy of insurance was void.

We do not see such a provision as being something which would be opposed by insurers. Instead insurers would engage with auditors and engineers to ensure that effective D&O insurance coverage would be easily available for companies who prepared pre-written modern Seven Statement depositions and affidavits and held these in draft in their business records, ready for use as and when required.

Alistair Kelman and James Christie

20 March 2025