PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Cite this article: Gunn LJ, Chapeau-Blondeau F, McDonnell MD, Davis BR, Allison A, Abbott D. 2016 Too good to be true: when overwhelming evidence fails to convince. *Proc. R. Soc. A* **472**: 20150748. http://dx.doi.org/10.1098/rspa.2015.0748

Received: 28 October 2015
Accepted: 19 February 2016

Subject Areas:
cryptography, systems theory, statistics

Keywords:
Bayesian, cryptography, criminology

Author for correspondence:
Lachlan J. Gunn
e-mail: lachlan.gunn@adelaide.edu.au

Too good to be true: when overwhelming evidence fails to convince

Lachlan J. Gunn[1], François Chapeau-Blondeau[2], Mark D. McDonnell[1,3], Bruce R. Davis[1], Andrew Allison[1] and Derek Abbott[1]

[1]School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide 5005, Australia
[2]Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), University of Angers, 62 avenue Notre Dame du Lac, Angers 49000, France
[3]School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, South Australia 5095, Australia

LJG, 0000-0003-1767-7897

Is it possible for a large sequence of measurements or observations, which support a hypothesis, to counterintuitively decrease our confidence? Can unanimous support be too good to be true? The assumption of independence is often made in good faith; however, rarely is consideration given to whether a systemic failure has occurred. Taking this into account can cause certainty in a hypothesis to decrease as the evidence for it becomes apparently stronger. We perform a probabilistic Bayesian analysis of this effect with examples based on (i) archaeological evidence, (ii) weighing of legal evidence and (iii) cryptographic primality testing. In this paper, we investigate the effects of small error rates in a set of measurements or observations. We find that even with very low systemic failure rates, high confidence is surprisingly difficult to achieve; in particular, we find that certain analyses of cryptographically important numerical tests are highly optimistic, underestimating their false-negative rate by as much as a factor of 2^{80}.

1. Introduction

In a number of branches of science, it is now well known that deleterious effects can conspire to produce

THE ROYAL SOCIETY PUBLISHING

a benefit or desired positive outcome. A key example where this manifests is in the field of *stochastic resonance* [1–3], where a small amount of random noise can surprisingly improve system performance, provided some aspect of the system is nonlinear. Another celebrated example is that of *Parrondo's Paradox*, where individually losing strategies combine to provide a winning outcome [4,5].

Loosely speaking, a small amount of 'bad' can produce a 'good' outcome. But is the converse possible? Can too much 'good' produce a 'bad' outcome? In other words, can we have too much of a good thing?

The answer is affirmative—when improvements are made that result in a worse overall outcome this situation is known as *Verschlimmbesserung* [6] or *disimprovement*. While this converse paradigm is less well known in the literature, a key example is the *Braess Paradox*, where an attempt to improve traffic flow by adding bypass routes can counterintuitively result in worse traffic congestion [7–9]. Another example is the *truel*, where three gunmen fight to the death—it turns out that under certain conditions the weakest gunman surprisingly reduces his chances of survival by firing a shot at either of his opponents [10]. Similar effects have been shown to apply in social settings, with increasing success sometimes indicating a reduction in actual skill [11]. These phenomena can be broadly considered to fall under the class of anti-Parrondo effects [12,13], where the inclusion of winning strategies fail.

In this paper, for the first time, we perform a Bayesian mathematical analysis to explore the question of multiple confirmatory measurements or observations for showing when they can—surprisingly—disimprove confidence in the final outcome. We choose the striking example that increasing confirmatory identifications in a police *line-up* or *identity parade* can, under certain conditions, reduce our confidence that a perpetrator has been correctly identified.

Imagine that as a court case drags on, witness after witness is called. Let us suppose 13 witnesses have testified to having seen the defendant commit the crime. Witnesses may be notoriously unreliable, but the sheer magnitude of the testimony is apparently overwhelming. Anyone can make a misidentification but intuition tells us that, with each additional witness in agreement, the chance of them all being incorrect will approach zero. Thus, one might naively believe that the weight of as many as 13 unanimous confirmations leaves us beyond reasonable doubt.

However, this is not necessarily the case and more confirmations can surprisingly disimprove our confidence that the defendant has been correctly identified as the perpetrator. This type of possibility was recognized intuitively in ancient times. Under ancient Jewish law [14], one could not be unanimously convicted of a capital crime—it was held that the absence of even one dissenting opinion among the judges indicated that there must remain some form of undiscovered exculpatory evidence.

Such approaches are greatly at odds with standard practice in engineering, where measurements are often taken to be independent. When this is so, each new measurement tends to lend support to the outcome with which it most concords. An important question, then, is to distinguish between the two types of decision problem; those where additional measurements truly lend support, and those for which increasingly consistent evidence either fails to add or actively reduces confidence. Otherwise, it is only later when the results come under scrutiny that unexpectedly good results are questioned; Mendel's plant-breeding experiments provide a good example of this [15,16]; his results matching their predicted values sufficiently well that their authenticity has been mired in controversy since the early twentieth century.

The key ingredient is the presence of a hidden failure state that changes the measurement response. This change may be *a priori* quite rare—in the applications that we shall discuss, it ranges from 10^{-1} to 10^{-19}—but when several observations are aggregated, the *a posteriori* probability of the failure state can increase substantially, and even come to dominate the *a posteriori* estimate of the measurement response. We shall show that by including error rates, this changes the information-fusion rule in a measurement-dependent way. Simple linear superposition no longer holds, resulting in non-monotonicity that leads to these counterintuitive effects.

This paper is constructed as follows. First, we introduce an example of a hypothetical archaeological find: a clay pot from the Roman era. We consider multiple confirmatory measurements that decide whether the pot was made in Britain or Italy. Via a Bayesian analysis, we then show that due to failure states, our confidence in the pot's origin does not improve for large numbers of confirmatory measurements. We begin with this example of the pot, due to its simplicity and that it captures the essential features of the problem in a clear manner.

Second, we build on this initial analysis and extend it to the problem of the police identity parade, showing our confidence that a perpetrator has been identified surprisingly declines as the number of unanimous witnesses becomes large. We use this mathematical framework to revisit a specific point of ancient Jewish law—we show that it does indeed have a sound basis, even though it grossly challenges our naive expectation.

Third, we finish with a final example to show that our analysis has broader implications and can be applied to electronic systems of interest to engineers. We chose the example of a cryptographic system and that a surprisingly small bit error rate can result in a larger-than-expected reduction in security.

Our analyses, ranging from cryptography to criminology, provide examples of how rare failure modes can have a counterintuitive effect on the achievable level of confidence.

2. A hypothetical Roman pot

Let us begin with a simple scenario, the identification of the origin of a clay pot that has been dug from British soil. Its design identifies it as being from the Roman era, and all that remains is to determine whether it was made in Roman-occupied Britain or whether it was brought from Italy by travelling merchants. Suppose that we are fortunate and that a test is available to distinguish between the clay from the two regions; clay from one area—let us suppose that it is Britain—contains a trace element which can be detected by laboratory tests with an error rate $p_e = 0.3$. This is clearly excessive, and so we run the test several times. After k tests have been made on the pot, the number of errors will be binomially distributed $E \sim \mathrm{Bin}(k, p_e)$. If the two origins, Britain and Italy, are *a priori* equally likely, then the most probable origin is the one suggested by the greatest number of samples.

Now imagine that several manufacturers of pottery deliberately introduced large quantities of this element during their production process, and that therefore it will be detected with 90% probability in their pots, which make up $p_c = 1\%$ of those found; of these, half are of British origin. We call p_c the *contamination rate*. This is the hidden failure state to which we alluded in the Introduction. Then, after the pot tests positive several times, we will become increasingly certain that it was manufactured in Britain. However, as more and more test results are returned from the laboratory, all positive, it will become more and more likely that the pot was manufactured with this unusual process, eventually causing the probability of British origin, given the evidence, to fall to 50%. This is the essential paradox of the system with hidden failure states—overwhelming evidence can itself be evidence of uncertainty, and thus be less convincing than more ambiguous data.

(a) Formal model

Let us now proceed to formalize the problem above. Suppose we have two hypotheses, H_0 and H_1, and a series of measurements $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. We define a variable $F \in \mathbb{N}$ that determines the underlying measurement distribution, $p_{X|F,H_i}(x)$. We may then use Bayes' Law to find

$$P[H_i|\mathbf{X}] = \frac{P[\mathbf{X}|H_i]P[H_i]}{P[\mathbf{X}]}, \tag{2.1}$$

which can be expanded by condition with respect to F, yielding

$$= \frac{\sum_f P[\mathbf{X}|H_i, f]P[H_i, F = f]}{\sum_{f,H_k} P[\mathbf{X}|H_k, f]P[H_k, F = f]}. \tag{2.2}$$

Table 1. The model parameters for the case of the pot for use in equation (2.3) with a contamination rate $p_c = 10^{-2}$. The *a priori* distribution of the origin is identically 50% for both Britain and Italy, whether or not the pot's manufacturing process has contaminated the results. As a result, the two columns of $P[F, H_i]$ are identical. The columns of the measurement distribution, shown right, differ from one another, thereby giving the test discriminatory power. When the pot has been contaminated, the probability of a positive result is identical for both samples, rendering the test ineffective.

$P[F, H_i]$

		origin: Italy H_0	origin: Britain H_1
contaminated	Y $F=0$	0.005 $\frac{1}{2}p_c$	0.005 $\frac{1}{2}p_c$
	N $F=1$	0.495 $\frac{1}{2}(1-p_c)$	0.495 $\frac{1}{2}(1-p_c)$

each square represents an outcome

$P[\text{positive result} \mid F, H_i]$

		origin: Italy H_0	origin: Britain H_1
contaminated	Y $F=0$	0.9	0.9
	N $F=1$	0.3 p_e	0.7 $1-p_e$

each square represents a distribution

In our examples, there are a number of simplifying conditions—there are only two hypotheses and two measurement distributions, reducing equation (2.2) to

$$P[H_i|\mathbf{X}] = \left(1 + \frac{\sum_{f=0}^{1} P[\mathbf{X}|H_{1-i},\ F=f]P[H_{1-i}, F=f]}{\sum_{f=0}^{1} P[\mathbf{X}|H_i,\ F=f]P[H_i,\ F=f]}\right)^{-1}. \tag{2.3}$$

Computation of these *a posteriori* probabilities thus requires knowledge of two distributions: the measurement distributions $P[\mathbf{X}|H_k, F]$, and the state probabilities $P[H_i, F]$. Having tabulated these, we may substitute them into equation (2.3), yielding the *a posteriori* probability for each hypothesis. In this paper, the measurement distributions $P[\mathbf{X}|H_i, F=f]$ are all binomial; however, this is not the case in general.

(b) Analysis of the pot origin distribution

In the case of the pot, the hypotheses and measurement distributions—the origin and contamination, respectively—are shown in table 1.

Each measurement is Bernoulli-distributed, and the number of positive results is therefore described by a Binomial distribution, with the probability mass function

$$P[X=x] = \binom{N}{x} p^x (1-p)^{N-x}$$

after N trials, the probability p being taken from the measurement distribution section of table 1.

Substituting these probability masses into equation (2.2), we see in figure 1 that as more and more tests return positive results, we become increasingly certain of its British heritage, but an unreasonably large number of positive results will indicate contamination and so yield a reduced level of certainty.

It is worth taking a moment, however, to briefly discuss the effects of weakening certain conditions; in particular, we consider two cases: that where the rate of contamination depends upon the origin of the pot, and that where the results after contamination are also origin-dependent.

Where the rate of contamination depends upon the origin, evidence of contamination provides some small evidence of where the pot came from. This corresponds to the case where the two cells in the top-most row of the left side of table 1 are different. For example, if 80% of contaminated

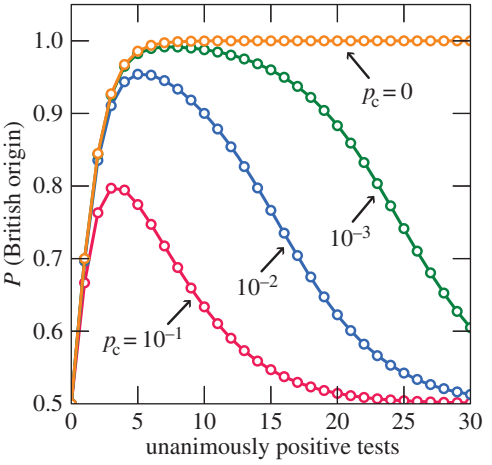

Figure 1. Probability that the pot is of British origin given N tests, all coming back positive, for a variety of contamination rates p_c and a 30% error rate. In the case of the pot above, with $p_c = 10^{-2}$, we see a peak at $N = 5$, after which the level of certainty falls back to 0.5 as it becomes more likely that the pot originates at a contaminating factory. When $p_c = 0$, this is the standard Bayesian analysis where failure states are not considered. We see therefore that even small contamination rates can have a large effect on the global behaviour of the testing methodology. (Online version in colour.)

pots are of British origin, then the as the number of unanimous results increases, the posterior distribution calculated in equation (2.2) will eventually converge to 0.8 rather than 0.5 as occurred in figure 1.

If the probability of a positive test is still slightly greater in the case of a British origin, even when contaminated—that is to say, when contamination merely reduces the accuracy of the test rather than rendering it useless—then the behaviour of the test protocol changes qualitatively. This corresponds to the case where the two cells in the top-most row of the right side of table 1 are different. Thus, as we receive more and more positive results, the probability of British origin as given by equation (2.2) will drop towards 0.5, the prior, due to the increasing likelihood of contamination; there might be enough positive results to determine that contamination has taken place, but not enough to conclude that it is of British origin. Eventually, however, it will rise again towards 1.0 as sufficient data becomes available to make use of the less probative experiment—the test in the contaminated case—that we now know to be taking place.

3. The reliability of identity parades

We initially described the scenario of a court case, in which witness after witness testifies to having seen the defendant commit the crime of which he is accused. But in-court identifications are considered unreliable, and in reality if identity is in dispute then the identification is made early in the investigation under controlled conditions [17]. The role of a police investigation is to find the guilty party, the perpetrator of the crime. In doing so, they will find one or more suspects who may or may not be the perpetrator.

If the identity of the perpetrator is in question, then at some point the suspect will be shown separately to each witness, surrounded among a number of other people, known as fillers, who are not under suspicion. Each witness is asked to identify the true perpetrator, if present, among the group.

This process, known as an *identity parade* or *line-up*, can be treated as an experiment intended to determine whether the suspect is in fact the same person as the perpetrator, a fact which is not known by the police, who take the role of the experimenter. Each witness that independently selects the suspect is an observation in support of that hypothesis. It may be performed only

once, or repeated many times with many witnesses. As human memory is inherently uncertain, the process will include random error; if the experiment is not properly carried out then there may also be systematic error, and this is the problem that concerns us in this paper.

Having seen how a unanimity of evidence can create uncertainty in the case of the unidentified pot, we now apply the same analysis to the case of an identity parade. If the perpetrator is not present—that is to say, if the suspect is innocent—then in an unbiased parade no witness should be able to choose the suspect with a probability greater than chance. Ideally, they would decline to make a selection; however, this does not always occur in practice [17,18], and forms part of the random error of the procedure. If the parade is biased—whether intentionally or unintentionally—for example, because (i) the suspect is somehow conspicuous [19], (ii) the staff running the parade direct the witnesses towards him, (iii) by chance he happens to resemble the perpetrator more closely than the fillers, or (iv) because the witness holds a bias, for example because they have previously seen the suspect [17], then an innocent suspect may be selected with a probability greater than chance. This is the hidden failure state that underlies this example; we assume in our analysis that this is completely binary—either the parade is completely unbiased or it is highly biased against the suspect.

In recent decades, a number of experiments [18,20] have been carried out in order to establish the reliability of this process. Test subjects are shown the commission of a simulated crime, whether in person or on video, and asked to locate the perpetrator among a number of people. In some cases, the perpetrator will be present, and in others not. The former allows estimation of the false-negative rate of the process—the rate that the witness fails to identify the perpetrator when present—and the latter the false-positive rate—the rate at which an innocent suspect will be mistakenly identified. Let us denote by p_{fn} the false-negative rate; this is equal to the proportion of subjects who failed to correctly identify the perpetrator when he was present and was found in [18] to be 48%.

Estimating the false-positive rate requires that we consider the number of others, the fillers, present in the line-up—when the suspect is innocent, an eyewitness who incorrectly identifies a filler as being the perpetrator has correctly rejected the innocent suspect as being the perpetrator, despite their error. For the purposes of our analysis, we assume that each witness selects at random if the suspect were not at the crime scene.

To find the probability that an innocent suspect will be selected in an unbiased line-up, we therefore divide the 80% perpetrator-absent selection rate from [18] by the number of participants $L = 6$, yielding a false-positive rate of $p_{\mathrm{fp}} = 0.133$.

Let us now suppose that there is a small probability p_{c} that the line-up is conducted incorrectly—for example, volunteers have been chosen who fail to adequately match the description of the perpetrator—leading the witness to point to the suspect 90% of the time, irrespective of his guilt. For the sake of analysis, we assume that if this occurs, it will occur for all witnesses, though in practice the police might perform the procedure correctly for some witnesses and not others. The probability of the suspect being identified for each case is shown in table 2.

If we assume a 50% prior probability of guilt, and independent witnesses, the problem is now identical to that of identifying the pot. The posterior probability of guilt, given the unanimous parade results, is given by equation (2.3) and shown in figure 2 as a function of the number of unanimous witnesses.

We see that after a certain number of unanimously positive identifications the posterior probability of guilt diminishes. Even with only one in 10 000 line-ups exhibiting this bias towards the suspect, the peak posterior probability is reached with only five unanimous witnesses, completely counter to intuition—in fact, with this rate of failure, 10 identifications in agreement provide less evidence of guilt than three. We see also that even with a 50% prior probability of guilt, a 1% failure rate renders it impossible to achieve 95% certainty if the witnesses are unanimous.

This tendency to be biased towards a particular member of the line-up when an error occurs has been noted [17, paragraph 4.31] prior to the more rigorous research stimulated by the advent

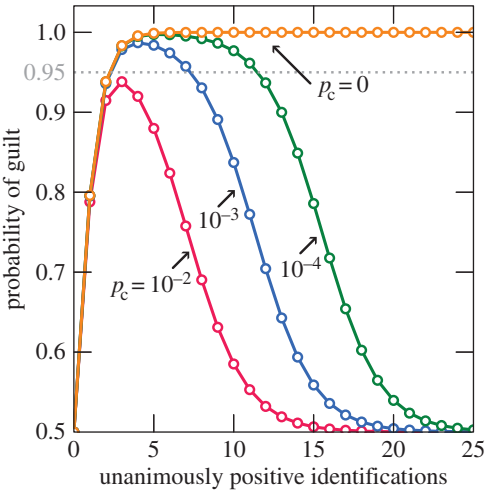

Figure 2. Probability of guilt given varying numbers of unanimous line-up identifications, assuming a 50% prior probability of guilt and identification accuracies given by Foster *et al*. [18]. Of note is that for the case that we have plotted here where the witnesses are unanimous, with a failure rate $p_c = 0.01$ it is impossible to reach 95% certainty in the guilt of the suspect, no matter how many witnesses have been found. (Online version in colour.)

Table 2. The model parameters for the hypothetical identity parade. In a similar manner to the first example, we assume *a priori* a 50% probability of guilt. In this case, the measurement distributions are substantially asymmetric with respect to innocence and guilt, unlike table 1 in which the lower row of the measurement distribution sums to one. This means that the well-performed identity parade is biased towards ruling out the suspect.

$P[F, H_i]$

biased parade	suspect is... innocent H_0	suspect is... guilty H_1
Y $F = 0$	0.005 $\frac{1}{2} p_c$	0.005 $\frac{1}{2} p_c$
N $F = 1$	0.495 $\frac{1}{2}(1 - p_c)$	0.495 $\frac{1}{2}(1 - p_c)$

each square represents an outcome

$P[\text{identification} \mid F, H_i]$

biased parade	suspect is... innocent H_0	suspect is... guilty H_1
Y $F = 0$	0.9	0.9
N $F = 1$	0.13 p_{fp}	0.52 $1 - p_{fn}$

each square represents a distribution

of DNA testing, leading us to suspect that our sub-1% contamination rates are probably overly optimistic.

4. Ancient judicial procedure

The acknowledgement of this type of phenomenon is not entirely new; indeed, the adage 'too good to be true' dates to the sixteenth century [21, *good*, P5.b]. Moreover, its influence on judicial procedure was visible in Jewish law even in the classical era; until the Romans ultimately removed the right of the Sanhedrin to confer death sentences, a defendant unanimously condemned by the judges would be acquitted [14, Sanhedrin 17a], the Talmud stating 'If the Sanhedrin unanimously find guilty, he is acquitted. Why? — Because we have learned by tradition that sentence must be postponed till the morrow in hope of finding new points in favour of the defence'.

Table 3. The model parameters for the Sanhedrin trial. Again, we assume an *a priori* 50% probability of guilt. However, the measurement distributions are the results of [22, model (2)] for juries; in contrast to the case of the identity parade, the false-negative rate is far lower. Despite the trial being conducted by judges, we choose to use the jury results, as the judges tendency towards conviction is not reflected in the highly risk-averse rabbinic legal tradition.

$P[F, H_i]$

biased proceedings \ suspect is...	innocent H_0	guilty H_1
Y $F=0$	0.005 $\frac{1}{2}p_c$	0.005 $\frac{1}{2}p_c$
N $F=1$	0.495 $\frac{1}{2}(1-p_c)$	0.495 $\frac{1}{2}(1-p_c)$

each square represents an outcome

$P[\text{identification} \mid F, H_i]$

biased proceedings \ suspect is...	innocent H_0	guilty H_1
Y $F=0$	0.95	0.95
N $F=1$	0.14 p_{fp}	0.75 $1-p_{fn}$

each square represents a distribution

The value of this rule becomes apparent when we consider that the Sanhedrin was composed, for ordinary capital offences, of 23 members [14, Sanhedrin 2a]. In our line-up model, this many unanimous witnesses would indicate a probability of guilt scarcely better than chance, suggesting that the inclusion of this rule should have a substantial effect.

We show the model parameters for the Sanhedrin decision in table 3, which we use to compute the probability of guilt using equation (2.3). The *a posteriori* probability of guilt is shown in figure 3 for various numbers of judges condemning the defendant. We see that the probability of guilt falls as judges approach unanimity; however, excluding unanimous decisions substantially reduces the probability of false conviction.

It is worth stressing that the exact shapes of the curves in figure 3 are unlikely to be entirely correct; communication between the judges will prevent their verdicts from being entirely independent, and false-positive and false-negative rates will be very much dependent upon the evidentiary standard required to bring charges, the strength of the contamination when it does occur, and the accepted burden of proof of the day. However, it is nonetheless of qualitative interest that with reasonable parameters, this ancient law can be shown to have a sound statistical basis.

5. The reliability of cryptographic systems

We now consider a different example, drawn from cryptography. An important operation in many protocols is the generation and verification of prime numbers; the security of some protocols depends upon the primality of a number that may be chosen by an adversary; in this case, one may test whether it is a prime, whether by brute-force or by using another test such as the Rabin–Miller [23, p. 176] test. As the latter is probabilistic, we repeat it until we have achieved the desired level of security—in [23], a probability 2^{-128} of accepting a composite as prime is considered acceptable. However, a naive implementation cannot achieve this level of security, as we will demonstrate.

The reason is that despite it being proven that each iteration of the Rabin–Miller test will reject a composite number with probability at least 0.75, a real computer may fail at any time. The chance of this occurring is small; however, it turns out that the probability of a stray cosmic ray flipping a bit in the machine code, causing the test to accept composite numbers, is substantially greater than 2^{-128}.

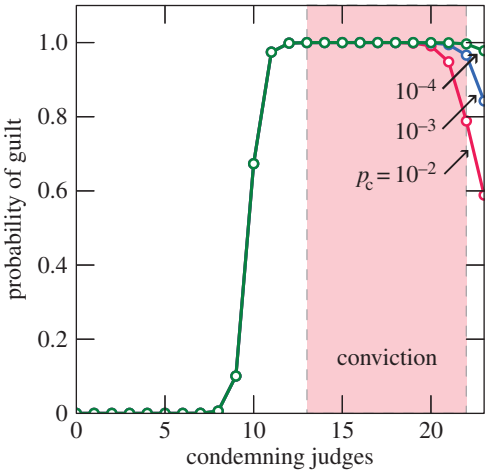

Figure 3. Probability of guilt as a function of judges in agreement out of 23—the number used by the Sanhedrin for most capital crimes—for various contamination rates p_c. We assume as before that half of defendants are guilty, and use the estimated false-positive and false-negative rates of juries from [22, model (2)], 0.14 and 0.25, respectively. We arbitrarily assume that a 'contaminated' trial will result in the a positive vote 95% of the time. The panel of judges numbers 23, with conviction requiring a majority of two and at least one dissenting opinion [14, Sanhedrin]; the majority of two means that the agreement of at least 13 judges is required in order to cast a sentence of death, to a maximum of 22 votes in order to satisfy the requirement of a dissenting opinion. These necessary conditions for a conviction by the Sanhedrin are shown as the pink region in the graph. (Online version in colour.)

(a) Code changes caused by memory errors

Data provided by Google [24] suggest that a given memory module has approximately an 8% probability of suffering an error in any given year, independent of capacity. Assuming a 4GB module, this results in approximately a $\lambda = 10^{-19}$ probability that any given bit will be flipped in any given second. We will make the assumption that, in the machine code for the primality-testing routine, there exists at least one bit that, if flipped, will cause all composite numbers—or some class of composite numbers known to the adversary—to be accepted as prime. As an example of how this could happen, consider the function shown in figure 4 that implements a brute-force factoring test. Assuming that the input is odd, the function will reach one of two return statements, returning zero or one. The C compiler GCC compiles these two return statements to

```
 45 0053 B8010000          movl     $1, %eax
 45      00
 46 0058 EB14              jmp    .L3
```

and

```
 56 0069 B8000000          movl $0, %eax
 56      00
 57                .L3:
```

respectively. That is to say, it stores the return value as an immediate into the EAX register and then jumps to the cleanup section of the function, labelled `.L3`. The store instructions on lines 45 and 56 have machine-code values `B801000000` and `B8000000000` for return values of one and zero, respectively. These differ by only one bit, and therefore can be transformed into one another by a single bit-error. If the first instruction is turned into the second, this will cause the function to return zero for any odd input, thus always indicating that the input is prime.

```
int  trialdivision (long to_test)
{
        long i;
        long threshold;

        if( to_test  % 2 == 0)
        {
                return 1;
        }

        threshold = (long)sqrt( to_test );

        for(i  = 3;  i  <= threshold;  i  += 2)
        {
                if( to_test  % i  == 0)
                {
                        return 1;
                }
        }

        return 0;
}
```

Figure 4. A function that tests for primality by attempting to factorize its input by brute force.

(b) The effect of memory errors on confidence

At cryptographically interesting sizes—of the order of 2^{2000}—roughly one in a 1000 numbers is prime [23, p. 173]. We might calculate the model parameters as before—for the sake of interest, we have done so in table 4—and calculate the confidence in a number's primality after a given number of tests. However, this is not particularly useful, for two reasons: first, the rejection probability of 75% is a lower bound, and for randomly chosen numbers is a substantial underestimate; second, we do not always choose numbers at random, but rather may need to test those provided by an adversary. In this case, we must assume that they have tried to deceive us by providing a composite number, and would instead like to know the probability that they will be successful. The Bayesian estimator in this case would provide only a tautology of the type: 'given the data and the fact that the number is composite, the number is composite'.

Let us suppose that the machine containing the code is rebooted every month, and the Rabin–Miller code remains in memory for the duration of this period; then, neglecting other potential errors that could affect the test, at the time of the reboot the probability that the bit has flipped is now $p_f = 2.6 \times 10^{-13}$; this event we denote by A_F. Let k be the number of iterations performed; the probability of accepting a composite number is at most 4^{-k}, and we assume that the adversary has chosen a composite number such that this is the true probability of acceptance. We denote by the event that the prime is accepted by the correctly operating algorithm A_R.

When hardware errors are taken into account, the probability of accepting a composite number is no longer 4^{-k}, but

$$p_{\text{fa}} = P[A_F \cup A_R] \tag{5.1}$$

$$= P[A_F] + P[A_R] - P[A_F, A_R]. \tag{5.2}$$

Since A_F and A_R are independent,

$$= P[A_F] + P[A_R] - P[A_F]P[A_R] \tag{5.3}$$

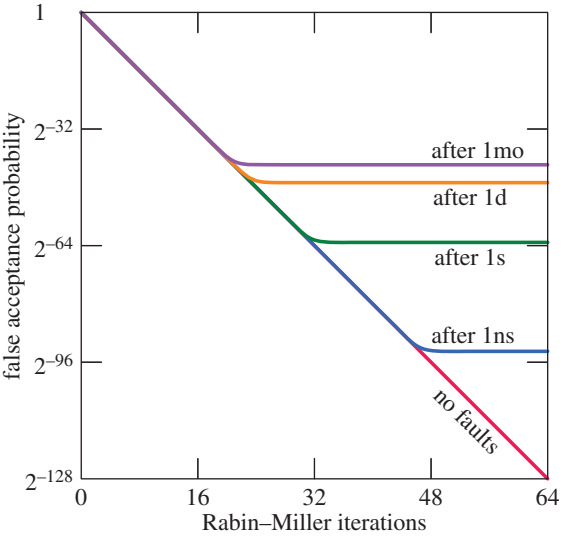

Figure 5. The acceptance rate as a function of time in memory and the number of Rabin–Miller iterations under the single-error fault model described in this paper. An acceptance rate of 2^{-128} is normally chosen, however, without error correction this cannot be achieved. The false-acceptance rate after k iterations is given by $p_{\mathrm{fa}}[k] = 4^{-k}(1 - p_{\mathrm{f}}) + p_{\mathrm{f}}$, where p_{f} is the probability that a fault has occurred that causes a false acceptance 100% of the time. We estimate p_f to be equal to $10^{-19}T$, where T is the length of time in seconds that the code has been in memory. (Online version in colour.)

Table 4. Model parameters for the Rabin–Miller test on random 2000-bit numbers. However, we have no choice but to assume the lower bound on the composite-number rejection rate, and so this model is inappropriate. Furthermore, in an adversarial setting the attacker may intentionally choose a difficult-to-detect composite number, rendering the prior distribution optimistic.

$P[F, H_i]$

always positive result	number is... prime H_0	number is... composite H_1
Y $F=0$	0.001×10^{-13} $0.001p_c$	0.999×10^{-13} $0.999\,p_c$
N $F=1$	≈0.001 $0.001\,(1-p_c)$	≈0.999 $0.999\,(1-p_c)$

each square represents an outcome

$P[\text{acceptance} \mid F, H_i]$

always positive result	number is... prime H_0	number is... composite H_1
Y $F=0$	1.0	1.0
N $F=1$	1.0	0.25

each square represents a distribution

$$= 4^{-k}(1 - p_f) + p_f \tag{5.4}$$

$$\geq p_f. \tag{5.5}$$

No matter how many iterations k of the algorithm are performed, this is substantially greater than the 2^{-128} security level that is predicted by probabilistic analysis of the algorithm alone, thus demonstrating that algorithmic analyses that do not take into account the reliability of the underlying hardware can be highly optimistic. The false acceptance rate as a function of the number of test iterations and time in memory is shown in figure 5.

A real cryptographic system will include many such checks in order to make sure that an attacker has not chosen weak values for various parameters, and a failure of any of these may result in the system being broken, so our calculations are somewhat optimistic.

Error-correcting-code equipped memory will substantially reduce the risk of this type of fault, and for regularly accessed regions of code—multiple times per second—will approach the 2^{-128} level. A single parity bit, as used in at least some CPU level-one instruction caches [25], requires two bit-flips to induce an error. Suppose the parity is checked every R seconds, then the probability of an undetected bit-flip in any given second is

$$\lambda' = \frac{(\lambda R)^2}{R} = \lambda^2 R. \tag{5.6}$$

For code that is accessed even moderately often, this will come much closer to 2^{-128}. For example, if $R = 100$ ms then this results in a false-acceptance rate of 2^{-108} after one month, much closer to the 2^{-128} level of security promised by analysis of the algorithm alone. The stronger error-correction codes used by the higher level caches and main memory will detect virtually all such errors—with two-bit detection capability, the rate of undetected bit-flips will be at most

$$\lambda' = \lambda^3 R^2, \tag{5.7}$$

and even with check rate of only once per 100 ms, the rate of memory errors is essentially zero, increasing the false-acceptance rate by a factor of only 10^{-14} above the 2^{-128} level that would be achieved in a perfect operating environment.

6. Discussion

This phenomenon is interesting in that it is commonly known and applied heuristically, and trivial examples such as the estimation of coin bias [26, section 2.1] have been well analysed—see appendix A for a brief discussion—but these uncommon failure states are rarely, if ever, considered when a statistical approach to decision-making is applied to an entire system. Real systems that attempt to counter failure modes producing consistent data tend to focus upon the detection of particular failures rather than the mere fact of consistency. Sometimes, there is little choice—a casino that consistently ejected gamblers on a winning streak would soon find itself without a clientelle—however, we have demonstrated that in many cases the level of consistency needed to cast doubt on the validity of the data is surprisingly low.

If this is so, then we must reconsider the use of thresholding as a decision mechanism when there is the potential for such failure modes to exist, particularly when the consequences of an incorrect decision are large. When the decision rule takes the form of a probability threshold, it is necessary to deduce an upper threshold as well, such as was shown in figure 3, in order to avoid capturing the region indicative of a systemic failure.

That this phenomenon was accounted for in ancient Jewish legal practice indicates a surprising level of intuitive statistical sophistication in this ancient law code; though predating by millennia the statistical tools needed to perform a rigorous analysis, our simple model of the judicial panel indicates that the requirement of a dissenting opinion would have provided a substantial increase in the probability of guilt required to secure a conviction.

Applied to cryptographic systems, we see that even the minuscule probability that one particular bit in the system's machine code will be flipped due to a memory error over the course of a month, rendering the system insecure, is approximately 2^{80} times larger than the risk predicted by algorithmic analysis. This demonstrates the importance of strong error correction in modern cryptographic systems that strive for a failure rate of the order of 2^{-128}, a level of certainty that appears to be otherwise unachievable without active mitigation of the effect.

The use of naturally occurring memory errors for DNS hijacking [27] has previously been demonstrated, and the ability of a user to disturb protected addresses by writing to adjacent cells [28] has been demonstrated; however, little consideration has been given to the possibility that this type of fault might occur simply by chance, implying that security analyses which

assume reliable hardware are substantially flawed when applied to consumer systems lacking error-corrected memory.

We have considered only a relatively simple case, in which there are only two levels of contamination. However, in practical situations we might expect any of a wide range of failure modes varying continuously. We have described a simple case in appendix A, where a coin may be biased—or not—towards either heads or tails with any strength; were one to apply this to the case of an identity parade, for example, one would find a probability that the suspect is indeed the perpetrator, as before, but taking into account that there may well be slight biases that nudge the witnesses towards or away from the suspect, not merely catastrophic ones. The result is heavily dependent upon the distribution of the bias, and lacking sufficient data to produce such a model we have chosen to eschew the complexity of the continuous approach and focus on a simple two-level model. An example of this approach is shown in [29, p. 117].

A related concept to that which we have discussed is the Duhem–Quine hypothesis [29, p. 6]; this is the idea that an experiment inherently tests hypotheses as a group—not merely the phenomenon that we wish to examine, but also the correct function of the experimental apparatus, for example, and that only the desired independent variables are being changed. Our thesis is a related one, namely that in practical systems the failure of these auxiliary hypotheses, though unlikely, result in a significant reduction in confidence when it occurs, an effect which has traditionally been ignored.

7. Conclusion

We have analysed the behaviour of systems that are subject to systematic failure, and demonstrated that with relatively low error rates, large sample sizes are not required in order that unanimous results start to become indicative of systematic failure. We have investigated the effect of this phenomenon upon identity parades, and shown that even with only a 1% rate of failure, confidence begins to decrease after only three unanimous identifications, failing to reach even 95%.

We have also applied our analysis of the phenomenon to cryptographic systems, investigating the effect by which confidence in the security of a parameter fails to increase with further testing due to potential failures of the underlying hardware. Even with a minuscule soft error rate of 10^{-13} per month, this effect dominates the analysis and is thus a significant determining factor in the overall level of security, increasing the probability that a maliciously chosen parameter will be accepted by a factor of more than 2^{80}.

Hidden failure states such as these reduce confidence far more than intuition leads one to believe, and must be more carefully considered than is the case today if the lofty targets that we set for ourselves are to be achieved in practice.

Data accessibility. This paper is mathematical, we have not generated new data.

Authors' contributions. L.J.G. drafted the manuscript. L.J.G. and D.A. devised the concept. L.J.G., F.C.-B., M.D.M., B.R.D., A.A. and D.A. carried out analyses and checking. All authors contributed to proofing the manuscript. All authors gave final approval for publication.

Competing interests. We have no competing interests.

Funding. L.J.G. is a Visiting Scholar at the University of Angers, France, supported by an Endeavour Research Fellowship from the Australian Government. M.D.M. is supported by an Australian Research Fellowship (DP1093425) and D.A. is supported by a Future Fellowship (FT120100351), both from the Australian Research Council (ARC).

Acknowledgements. We would like to thank Leila Schneps and Robert E. Bogner for critical comments on the draft manuscript.

Appendix A. Analysis of a biased coin

It is worth adding a brief discussion of a simple and well-known problem that has some relation to what we have discussed, namely the question of whether or not a coin is biased. We follow the Bayesian approach given in [26].

They use Bayes' law in its proportional form,

$$P[Q|\{\text{data}\}] \propto P[\{\text{data}\}|Q]P[Q], \tag{A 1}$$

where Q is the probability that a coin toss will yield heads. Various prior distributions $P[Q]$ can be chosen, a matter that we will discuss in a moment.

As the coin tosses are independent, the data can be boiled down to a binomial random variable $X \sim \text{Bin}(N, p)$, where N is the number of coin tosses made. Substituting the binomial probability mass function into equation (A 1), they find that

$$P[Q|X] \propto Q^X(1-Q)^{N-X}P[Q]. \tag{A 2}$$

As the number of samples N increases, this becomes increasingly peaked around the value $Q = X/N$; if the prior distribution $P[Q]$ assigns a low probability to this value, then it may take quite some time for the 'peaking' effect to manifest itself in the posterior distribution. Nonetheless, as the number of samples increases, the $Q^X(1-Q)^{N-X}$ part of the expression eventually comes to dominate the shape of the posterior distribution $P[Q|X]$, and we have no choice but to believe that the coin genuinely does have a bias close to X/N.

In the examples previously discussed, we have assumed that bias is very unlikely; in the coin example, this corresponds to a prior distribution $P[Q]$ that is strongly clustered around $Q = 0.5$; in this case, a very large number of samples will be necessary in order to conclusively reject the hypothesis that the coin is unbiased or nearly so. However, eventually this will occur, and the posterior distribution will change; when this occurs, the system has visibly failed—a casino using the coin will decide that they are not in fact playing the game that they had planned, and must cease before their loss becomes catastrophic. This is much like in the case of the Sanhedrin—if too many judges agree, the system has failed and should not be considered reliable.

References

1. Benzi R, Sutera A, Vulpiani A. 1981 The mechanism of stochastic resonance. *J. Phys. A Math. Gen.* **14**, L453–L457. (doi:10.1088/0305-4470/14/11/006)
2. McDonnell MD, Stocks NG, Pearce CEM, Abbott D. 2008 *Stochastic resonance: from suprathreshold stochastic resonance to stochastic signal quantization*. Cambridge, UK: Cambridge University Press.
3. McDonnell MD, Abbott D. 2009 What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology. *PLoS Comput. Biol.* **5**, e1000348. (doi:10.1371/journal.pcbi.1000348)
4. Harmer GP, Abbott D. 1999 Game theory: losing strategies can win by Parrondo's paradox. *Nature* **402**, 864. (doi:10.1038/47220)
5. Abbott D. 2010 Asymmetry and disorder: a decade of Parrondo's paradox. *Fluct. Noise Lett.* **9**, 129–156. (doi:10.1142/S0219477510000010)
6. Mead MN. 2005 Columbia program digs deeper into arsenic dilemma. *Environ. Health Perspect.* **113**, A374–A377. (doi:10.1289/ehp.113-a374)
7. Braess D. 1969 Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* **12**, 258–268.
8. Kameda H, Altman E, Kozawa T, Hosokawa Y. 2000 Braess-like paradoxes in distributed computer systems. *IEEE Trans. Autom. Control* **45**, 1687–1691. (doi:10.1109/9.880619)
9. Korilis YA, Lazar AA, Orda A. 1999 Avoiding the Braess paradox in noncooperative networks. *J. Appl. Probab.* **36**, 211–222. (doi:10.1239/jap/1032374242)
10. Flitney AP, Abbott D. 2004 Quantum two- and three-person duels. *J. Opt. B: Quant. Semiclass. Opt.* **6**, S860–S866. (doi:10.1088/1464-4266/6/8/036)
11. Denrell J, Liu C. 2012 Top performers are not the most impressive when extreme performance indicates unreliability. *Proc. Natl Acad. Sci. USA* **109**, 9331–9336. (doi:10.1073/pnas.1116048109)
12. Harmer GP, Abbott D, Taylor PG, Parrondo JMR. 2001 Brownian ratchets and Parrondo's games. *Chaos* **11**, 705–714. (doi:10.1063/1.1395623)
13. Ethier SN, Lee J. 2012 Parrondo's paradox via redistribution of wealth. *Electron. J. Probab.* **17**, 20. (doi:10.1214/EJP.v17-1867)

14. Epstein I (ed.). 1961 *The Babylonian Talmud*. London, UK: Soncino Press.
15. Fisher RA. 1936 Has Mendel's work been rediscovered? *Ann. Sci.* **1**, 115–137. (doi:10.1080/00033793600200111)
16. Franklin A. 2008 *Ending the Mendel-Fisher controversy*. Pittsburgh, PA: University of Pittsburgh Press.
17. Devlin P, Freeman C, Hutchinson J, Knights P. 1976 *Report to the secretary of state for the home department of the departmental committee on evidence of identification in criminal cases*. London, UK: HMSO.
18. Foster RA, Libkuman TM, Schooler JW, Loftus EF. 1994 Consequentiality and eyewitness person identification. *Appl. Cogn. Psychol.* **8**, 107–121. (doi:10.1002/acp.2350080203)
19. Wogalter MS, Marwitz DB. 1992 Suggestiveness in photospread lineups: similarity induces distinctiveness. *Appl. Cogn. Psychol.* **6**, 443–453. (doi:10.1002/acp.2350060508)
20. Malpass RS, Devine PG. 1981 Eyewitness identification: lineup instructions and the absence of the offender. *J. Appl. Psychol.* **66**, 482–489. (doi:10.1037/0021-9010.66.4.482)
21. Oxford English Dictionary. Oxford University Press, Oxford, UK. See http://oed.com/ (accessed 6 October 2015).
22. Spencer BD. 2007 Estimating the accuracy of jury verdicts. *J. Empir. Legal Stud.* **4**, 305–329. (doi:10.1111/j.1740-1461.2007.00090.x)
23. Ferguson N, Schneier B, Kohno T. 2010 *Cryptography engineering: design principles and practical applications*. Indianapolis, USA: Wiley.
24. Schroeder B, Pinheiro E, Weber W-D. 2009 DRAM errors in the wild: a large-scale field study. In *Proc. of the 11th Int. Joint Conf. on Measurement and Modeling of Computer Systems, SIGMETRICS '09, Seattle, WA, 15–19 June*, pp. 193–204. New York, NY: ACM.
25. Advanced Micro Devices. 2007 AMD Opteron processor product datasheet. See http://support.amd.com/TechDocs/23932.pdf (accessed 12 October 2015).
26. Sivia DS, Skilling J. 2006 *Data analysis: a Bayesian tutorial*. Oxford, UK: Oxford University Press.
27. Dinaberg A. 2011 Bitsquatting: DNS hijacking without exploitation. In *Proc. of BlackHat Security, Las Vegas, NV*. See https://media.blackhat.com/bh-us-11/Dinaburg/BH_US_11_Dinaburg_Bitsquatting_WP.pdf (accessed 8 March 2016).
28. Kim Y, Daly R, Kim J, Fallin C, Lee JH, Lee D, Wilkerson C, Lai K, Mutlu O. 2014 Flipping bits in memory without accessing them: an experimental study of DRAM disturbance errors. In *Proc. IEEE 41st Annual Int. Symp. on Computer Architecture, Minneapolis, MN, 13–18 June*, pp. 361–372. Piscataway, NJ: IEEE.
29. Bovens L, Hartmann S. 2004 *Bayesian epistemology*. Oxford, UK: Oxford University Press.