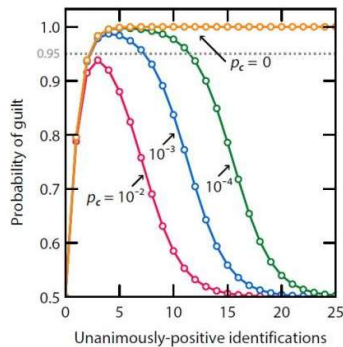


Why too much evidence can be a bad thing

4 January 2016, by Lisa Zyga



In a police line-up, the probability that an individual is guilty increases with the first three witnesses who unanimously identify him or her, but then decreases with additional unanimous witness identifications. Different colored lines represent various failure/error rates, with yellow representing zero failure. Credit: Gunn, et al. ©2016 The Royal Society

(Phys.org)—Under ancient Jewish law, if a suspect on trial was unanimously found guilty by all judges, then the suspect was acquitted. This reasoning sounds counterintuitive, but the legislators of the time had noticed that unanimous agreement often indicates the presence of systemic error in the judicial process, even if the exact nature of the error is yet to be discovered. They intuitively reasoned that when something seems too good to be true, most likely a mistake was made.

In a new paper to be published in *The Proceedings of The Royal Society A*, a team of researchers, Lachlan J. Gunn, et al., from Australia and France has further investigated this idea, which they call the "paradox of unanimity."

"If many independent witnesses unanimously testify to the identity of a suspect of a crime, we assume they cannot all be wrong," coauthor Derek Abbott, a physicist and electronic engineer at The University of Adelaide, Australia, told *Phys.org*. "Unanimity is often assumed to be reliable.

However, it turns out that the probability of a large number of people all agreeing is small, so our confidence in unanimity is ill-founded. This 'paradox of unanimity' shows that often we are far less certain than we think."

Unlikely agreement

The researchers demonstrated the paradox in the case of a modern-day police line-up, in which witnesses try to identify the suspect out of a line-up of several people. The researchers showed that, as the group of unanimously agreeing witnesses increases, the chance of them being correct decreases until it is no better than a random guess.

In police line-ups, the systemic error may be any kind of bias, such as how the line-up is presented to the witnesses or a personal bias held by the witnesses themselves. Importantly, the researchers showed that even a tiny bit of bias can have a very large impact on the results overall. Specifically, they show that when only 1% of the line-ups exhibit a bias toward a particular suspect, the probability that the witnesses are correct begins to decrease after only three unanimous identifications. Counterintuitively, if one of the many witnesses were to identify a different suspect, then the probability that the other witnesses were correct would substantially increase.

The mathematical reason for why this happens is found using Bayesian analysis, which can be understood in a simplistic way by looking at a biased coin. If a biased coin is designed to land on heads 55% of the time, then you would be able to tell after recording enough coin tosses that heads comes up more often than tails. The results would not indicate that the laws of probability for a binary system have changed, but that this particular system has failed. In a similar way, getting a large group of unanimous witnesses is so unlikely, according to the laws of probability, that it's more likely that the system is unreliable.

The researchers say that this paradox crops up more often than we might think. Large, unanimous agreement does remain a good thing in certain cases, but only when there is zero or near-zero bias. Abbott gives an example in which witnesses must identify an apple in a line-up of bananas—a task that is so easy, it is nearly impossible to get wrong, and therefore large, unanimous agreement becomes much more likely.

On the other hand, a criminal line-up is much more complicated than one with an apple among bananas. Experiments with simulated crimes have shown misidentification rates as high as 48% in cases where the witnesses see the perpetrator only briefly as he runs away from a crime scene. In these situations, it would be highly unlikely to find large, unanimous agreement. But in a situation where the witnesses had each been independently held hostage by the perpetrator at gunpoint for a month, the misidentification rate would be much lower than 48%, and so the magnitude of the effect would likely be closer to that of the banana line-up than the one with briefly seen criminals.

Wide implications

The paradox of unanimity has many other applications beyond the legal arena. One important one that the researchers discuss in their paper is cryptography. Data is often encrypted by verifying that some gigantic number provided by an adversary is prime or composite. One way to do this is to repeat a probabilistic test called the Rabin-Miller test until the probability that it mistakes a composite as prime is extremely low: a probability of 2^{-128} is typically considered acceptable.

The systemic failure that occurs in this situation is computer failure. Most people never consider the possibility that a stray cosmic ray may flip a bit that in turn causes the test to accept a composite number as a prime. After all, the probability for such an event occurring is extremely low, approximately 10^{-13} per month. But the important thing is that it's greater than 2^{-128} , so even though the failure rate is so tiny, it dominates over the desired level of security. Consequently, the cryptographic protocol may appear to be more secure than it really is, since test results that appear to indicate a high

level of security are actually much more likely to be indicative of computer failure. In order to truly achieve the desirable level of security, the researchers advise that these "hidden" errors must be reduced to as close to zero as possible.

The paradox of unanimity may be counterintuitive, but the researchers explain that it makes sense once we have complete information at our disposal.

"As with most 'paradoxes,' it is not that our intuition is necessarily bad, but that our intuition has been badly informed," Abbott said. "In these cases, we are surprised because we simply aren't generally aware that identification rates by witnesses are in fact so poor, and we aren't aware that bit error rates in computers are significant when it comes to cryptography."

The researchers noted that the paradox of unanimity is related to the Duhem-Quine hypothesis, which states that it is not possible to test a scientific hypothesis in isolation, but rather hypotheses are always tested as a group. For instance, an experiment tests not only a certain phenomenon, but also the correction function of the experimental tools. In the paradox of unanimity, it's the methods (the "auxiliary hypotheses") that fail, and in turn reduce confidence in the main results.

More examples

Other areas where the paradox of unanimity emerges are numerous and diverse. Abbott describes several below, in his own words:

1) The recent Volkswagen scandal is a good example. The company fraudulently programmed a computer chip to run the engine in a mode that minimized diesel fuel emissions during emission tests. But in reality, the emissions did not meet standards when the cars were running on the road. The low emissions were too consistent and 'too good to be true.' The emissions team that outed Volkswagen initially got suspicious when they found that emissions were almost at the same level whether a car was new or five years old! The consistency betrayed the systemic bias introduced by the nefarious computer chip.

- 2) A famous case where overwhelming evidence was 'too good to be true' occurred in the 1993-2008 period. Police in Europe found the same female DNA in about 15 crime scenes across France, Germany, and Austria. This mysterious killer was dubbed the Phantom of Heilbronn and the police never found her. The DNA evidence was consistent and overwhelming, yet it was wrong. It turned out to be a systemic error. The cotton swabs used to collect the DNA samples were accidentally contaminated, by the same lady, in the factory that made the swabs.
- 3) When a government wins an election, one laments that the party of one's choice often wins with a relatively small margin. We often wish for our favored political party to win with unanimous votes. However, should that ever happen we would be led to suspect a systemic bias caused by vote rigging. An urban legend persists that Putin won 140% (!) of the votes; if this true then democracy clearly failed in that case. The take-home message is that, in a healthy democracy, when a party wins by a small margin, instead of name-calling the 'dumb' voters of the opposition, we should be celebrating the fact that the opposing voters preserved the integrity of democracy.
- 4) In science, theory and experiment go hand in hand and must support each other. In every experiment there is always 'noise,' and we must therefore expect some error. In the history of science there are a number of famous experiments where the results were 'too good to be true.' There are many examples that have been mired in controversy over the years, and the most famous are Millikan's oil drop experiment for determining the charge on the electron and Mendel's plant breeding experiments. If results are too clean and do not contain expected noise and outliers, then we can be led to suspect a form of confirmation bias introduced by an experimenter who cherry-picks the data.
- 5) In many committee meetings, in today's big organizations, there is a trend towards the idea that decisions must be unanimous. For example, a committee that ranks job applicants or evaluates key performance indicators (KPIs) often will argue until everyone in the room is in agreement. If one or two members are in disagreement, there is a tendency for the rest of the committee to win them over before moving on. A take-home message of our analysis is that the dissenting voice should be welcomed. A wise committee should accept that difference of opinion and simply record there was a disagreement. The recording of the disagreement is not a negative, but a positive that demonstrates that a systemic bias is less likely.
- 6) Eugene Wigner once coined the phrase 'the unreasonable effectiveness of mathematics' to describe the rather odd feeling that math seems to be so perfectly suited to describing physical theories. In a way, Wigner was expressing the idea that math itself is 'too good to be true.' (See [this article](#) for more on this idea.) The reality is that modern devices and machines are no longer analyzed by neat analytical mathematical equations, but by empirical formulas embedded in simulation software tools. For some of the next big science questions, particularly in the area of complex systems, we are looking to big data and machine learning rather than math. Analytical math as we knew it was not the perfect glove that could fit every type of problem. So how did we get seduced to once thinking that math was 'unreasonably effective'? It's the systemic confirmation bias introduced by the fact that for every great scientific paper we read with an elegant formula, there are many more rejected formulas that are never published and we never get to see. The math we have today was cherry-picked.

More information: Lachlan J. Gunn, et al. "Too good to be true: when overwhelming evidence fails to convince." *Proceedings of The Royal Society A*. To be published. Arxiv pre-print: arxiv.org/abs/1601.00900

© 2016 Phys.org

APA citation: Why too much evidence can be a bad thing (2016, January 4) retrieved 12 January 2016 from <http://phys.org/news/2016-01-evidence-bad.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.