Trial by Jury

by Richard Dawkins. Published as "Three herring gull chicks . . . the reason juries don't work" in The Observer (London), Sunday November 16, 1997.

Trial by jury must be one of the most conspicuously bad good ideas anyone ever had. Its devisers can hardly be blamed. They lived before the principles of statistical sampling and experimental design had been worked out. They weren't scientists. Let me explain using an analogy. And if, at the end, somebody objects to my argument on the grounds that humans aren't herring gulls, I'll have failed to get my point across.

Adult herring gulls have a bright yellow bill with a conspicuous red spot near the tip. Their babies peck at the red spot, which induces the parents to regurgitate food for them. Niko Tinbergen, Nobel-Prizewinning zoologist and my old maestro at Oxford, offered naive young chicks a range of cardboard dummy gull heads varying in bill and spot colour, and shape. For each colour, shape or combination, Tinbergen measured the preferences of the baby chicks by counting their pecks in a standard time. The idea was to discover whether naive gull chicks are born with a built-in preference for long yellow things with red spots. If so, this would suggest that genes equip the young birds with detailed prior knowledge of the world in which they are about to hatch – a world in which food comes out of adult herring gull beaks.

Never mind the reason for the research, and never mind the conclusions. Consider, instead, the methods you must use, and the pitfalls you must avoid, if you want to get a correct result in any such experiment. These turn out to be general principles which apply to human juries as strongly as to gull chicks.

First, you obviously must test more than one chick. It could be that some chicks are red-biased, others blue-biased, with no tendency for herring gull chicks in general to share the same favourite colour. So, by picking out a single chick, you are measuring nothing more than individual bias. It is no answer to this objection that our chick may have given hundreds more pecks to one colour than to the other. A chick might begin by choosing any old colour at random, but once he has chosen he gets 'locked on' to that colour and hammers away at it, giving the other colours no chance. The essential problem here is that successive pecks, however numerous, are not 'independent data'.

So, we must test more than one chick. How many? Is two enough? No, nor is three, and now we must start to think statistically. To make it simple, suppose that in a particular experiment we are comparing only red spots versus blue spots, both on a yellow background, and always presented simultaneously. If we test just two chicks separately, suppose the first chick chooses red. It had a 50% chance of doing so, at random. Now the second chick also happens to choose red. Again, the odds were 50% that it would do so at random, even if it were colourblind. There's a 50% chance that two randomly choosing chicks will agree (half of the four possibilities: red red, red blue, blue red, blue blue). Three chicks aren't enough either. If you write down all the possibilities, you'll find that there's a 25% chance of a unanimous verdict, by luck alone. Twenty five percent, as the odds of reaching a conclusion for the wrong reason, is unacceptably large.

How about twelve good chicks and true? Now you're talking. If twelve chicks are independently offered a choice between two alternatives, the odds that they will all reach the same verdict by chance alone are satisfyingly low, only one in 1024.

But now suppose that, instead of testing our twelve chicks independently, we test them as a group. We take a maelstrom of twelve cheeping chicks and lower into their midst a red spotted dummy and a blue spotted dummy, each fitted with an electrical device for automatically tallying pecks. And suppose that the collective of chicks registers 532 pecks at red and zero at blue. Does this massive disparity show that herring gull chicks, in general, prefer red? Absolutely not. The pecks are not independent data. Chicks could have a strong tendency to imitate one another (as well as imitate themselves in lock-on effects). If one chick just happened to peck at red first, others might copy him

and the whole company of chicks join in a frenzy of imitative pecking. As a matter of fact this is precisely what domestic chicken chicks do, and gull chicks are very likely the same. Even if not, the principle remains that the data are not independent and the experiment is therefore invalid. The twelve chicks are strictly equivalent to a single chick, and their summed pecks amount to only a single independent result.

Turning to courts of law, why are twelve jurors preferred to a single judge? Not because they are wiser, more knowledgeable or more practised in the arts of reasoning. Certainly not, and with a vengeance. Think of the astronomical damages awarded by juries in footling libel cases. Think how juries bring out the worst in histrionic, gallery-playing lawyers. Twelve jurors are preferred to one judge only because they are more numerous. Letting a single judge decide a verdict would be like letting a single chick speak for the whole herring gull species. Twelve heads are better than one, because they represent twelve assessments of the evidence.

But for this argument to be valid, the twelve assessments really have to be independent. And of course they are not. Twelve men and women locked in a jury room are like our clutch of twelve gull chicks. Whether they actually imitate each other like chicks, they might. That is enough to invalidate the principle by which a jury might be preferred over a single judge.

In practice, as is well documented and as I remember from the three juries that it has been my misfortune to serve on, juries are massively swayed by one or two vocal individuals. There is also strong pressure to conform to a unanimous verdict, which further undermines the principle of independent data. Increasing the number of jurors doesn't help, or not much (and not at all in strict principle). What you have to increase is the number of independent verdict-reaching units.

Oddly enough, the bizarre American system of televising trials opens up a real possibility of improving the jury system. By the end of trials such as those of Louise Woodward or O.J.Simpson, literally thousands of people around the country have attended to the evidence as assiduously as the official jury. A mass phone-in might produce a fairer verdict than a jury. But unfortunately journalistic discussion, radio talk-shows, and ordinary gossip would violate the Principle of Independent Data and we'd be back where we started. The broadcasting of trials, in any case, has horrible consequences. In the wake of Louise Woodward's trial, the Internet seethes with ill-spelled and ungrammatical viciousness, the cheque-book journalists are queuing up, and the unfortunate Judge Zobel has had to change his telephone number and employ a bodyguard.

So, how can we improve the system? Should twelve jurors be locked in twelve isolation chambers and their opinions separately polled so that they constitute genuinely independent data? If it is objected that some would be too stupid or inarticulate to reach a verdict on their own, we are left wondering why such individuals are allowed on a jury at all. Perhaps there is something to be said for the collective wisdom that emerges when a group of twelve people thrash out a topic together, round a table. But this still leaves the principle of independent data unsatisfied.

Should all cases be tried by two separate juries? Or three? Or twelve? Too expensive, at least if each jury has twelve members. Two juries of six members, or three juries of four members, would probably be an improvement over the present system. But isn't there some way of testing the relative merits of such alternative options, or of comparing the merits of trial by jury versus trial by judge?

Yes, there is. I'll call it the Two Verdicts Concordance Test. It is based on the principle that, if a decision is valid, two independent shots at making it should yield the same result. Just for purposes of the test, we run to the expense of having two juries, listening to the same case and forbidden to talk to members of the other jury. At the end, we lock the two juries in two separate jury rooms and see if they reach the same verdict. If they don't, nothing can be proved beyond reasonable doubt, and this would cast reasonable doubt on the jury system itself.

To make the experimental comparison with Trial by Judge, we need two experienced judges to

listen to the same case, and require them too to reach their separate verdicts without talking to each other. Whichever system, Trial by Jury or Trial by Judge, yields the higher score of agreements over a number of trials is the better system and might even be accredited for future use with some confidence.

Would you bet on two independent juries reaching the same verdict in the Louise Woodward case? Could you imagine even one other jury reaching the same verdict in the O.J.Simpson case? Two judges, on the other hand, seem to me rather likely to score well on the concordance test. And should I be charged with a serious crime here's how I want to be tried. If I know myself to be guilty, I'll go with the loose cannon of a jury, the more ignorant, prejudiced and capricious the better. But if I am innocent, and the ideal of multiple independent decision-takers is unavailable, please give me a judge. Preferably Judge Hiller Zobel.