

Convergent evolution in law and science: the structure of decision-making under uncertainty

MICHAEL J. SAKS[†]

*Sandra Day O'Connor College of Law and Department of Psychology,
Arizona State University, 1100 S. McAllister, Tempe, AZ 85287-7906, USA*

AND

SAMANTHA L. NEUFELD

*Department of Psychology, Arizona State University, 950 S. McAllister,
Tempe, AZ 85287-1104, USA*

[Received on 4 November 2010; revised on 27 February 2011; accepted on 9 May 2011]

The formal structure of decision-making under uncertainty used in legal trials bears a noteworthy similarity to the structure of decision-making under uncertainty used in hypothesis testing in empirical science. The first purpose of this article was to explicate those similarities. Secondly, the article reviews the historical origins of these decision-making schemes in both law and science, finding that they evolved independent of each other to serve similar functions.

Keywords: decision-making; uncertainty; law and science; history of law and science; cultural evolution; convergent evolution.

1. Introduction

Discussions of the nature of law and science typically emphasize their differences.¹ The present article is an exception. We explicate a fundamental similarity of thought that evolved in these two different worlds when they confronted similar needs to deal with uncertainty. Indeed, to characterize these as 'similar' is to understate the parallels. In an area critical to each—decision-making under uncertainty—law and science have recognized, formalized and adopted approaches that are identical in most if not all important respects, though they use completely different language to describe the same things.

In addition, we investigate the historical origins of the two systems to see whether one informed the other or whether they evolved separately. Our review leads to the conclusion that they evolved completely independent of each other, much like species that evolve highly similar forms and structures in response to the same kinds of environmental challenges, though they have no biological

[†] Email: michael.saks@asu.edu

¹ For examples, see Cowan (1963), Faigman (1999), Goldberg (1996) and Haack (2009). Discussions of law and science often address such issues as the role of positive versus normative enquiry, central versus secondary role of truth-seeking, importance of established belief versus search for ideas that are better than what went before, seemingly irreconcilable tensions when they interact (regarding law's reliance on scientific knowledge as evidence as well as the legal regulation of scientific enquiry or application), and other issues.

connection to each other. This finding leads to our analogy to convergent evolution in legal trials and scientific hypothesis testing.

Before addressing the crux of the argument, we think that it is useful to define several terms that will appear repeatedly and to frame our paper in the wider context of decision theory. ‘Decision theory’, broadly, addresses the process of using known information to come to the best decision, given a set of options (Cox and Hinkley, 1974). Legal trials and scientific testing, along with a wide variety of other endeavors (medical screening, buying a home, etc.), fall under this umbrella term. Within decision theory is ‘statistical decision theory’, in which statistical analyses are conducted on the known information to reach a conclusion (Brown, 2000). Although law does not use statistics to make decisions, science relies quite extensively on these kinds of analyses. The foremost statistical testing procedure in science is ‘null hypothesis significance testing’ or NHST. The ‘null hypothesis’ typically states that the means associated with the two or more groups of interest are equal (i.e. the drug or manipulation has no effect on the dependent variable). In contrast, the ‘alternative hypothesis’ states that the means of the groups are not equal, with the implication that if they are not equal, it is because the experimental manipulation caused the difference (Cohen *et al.*, 2003; Keppel and Zedeck, 1989). NHST allows researchers to calculate the probability that they would have found what they found if the null hypothesis were true.

An important criticism of NHST is that it is an indirect test—calculating the probability that the finding would occur if the null were true is not the same as calculating the probability that the null is true, given the finding (Trafimow, 2003). This concern is addressed by ‘Bayesian theory’, which allows scientists to assess the probability that the null hypothesis is true, given the actual finding. Bayesian approaches emphasize a more informed starting point, thereby allowing more precise conclusions. As strong as the argument for Bayesian statistics is, however, it is often difficult to estimate the necessary prior probabilities (Trafimow, 2003). For that reason, NHST is still far more widely used in social, behavioural and biomedical science (although the push for Bayesian analysis is growing).

Two major topics follow. First, we review the concepts of decision-making under uncertainty commonly used in law and science. Then, we reflect on the history that brought courtroom trials and scientific hypothesis testing to their highly similar states.²

2. Decision-making under uncertainty

Each field is generally quite familiar with its own concepts of decision-making under uncertainty, but we suspect that few members of one field appreciate the strength of the parallels to be found in the other field. In the discussion below, we describe the two approaches in a way that highlights their similarities. Table 1 summarizes these shared elements, and the discussion below is keyed to that table.

² In a convergence that is both ironic and inevitable, numerous scholars of decision-making who approach the trial from a scientific perspective have placed the legal system’s version of what today is termed decision theory under various formal decision-theoretic microscopes. See e.g. Arkes and Mellers (2002), Connolly (1987), Kaplan (1968) and Thomas and Hogue (1976). In other words, if one drills down beneath the legal process to which these scholars have applied various aspects of decision theory, one finds a legal decision process that is virtually identical to a centrally important component of statistical decision theory. Some of the cited works apply as nicely as they do to the trial decision process precisely because that process already is well adapted to decision theory—indeed, in its own way, it anticipated decision theory. Neither these nor similar works of which we are aware make use of or take note of the parallelism that is the focus of the present article.

TABLE 1 *Parallel elements of hypothesis testing in law and science*

	Law	Science
1	Binary fallible decision: guilty or not guilty verdict (in criminal trials).	Binary fallible decision: reject or do not reject null hypothesis.
2	Two types of error: convict an innocent defendant or acquit a guilty one (in criminal trials).	Two types of error: reject a null hypothesis that is true or fail to reject a null hypothesis that is false.
3	A starting point that remains in force unless the evidence produced at trial justifies a different conclusion. That starting point is the presumption of innocence (in criminal trials).	A starting point that remains in force unless the evidence accumulated by the study justifies a different conclusion. That starting point is the null hypothesis.
4	Failing to reject the starting point is not equivalent to accepting the starting point as true.	Failing to reject the starting point is not equivalent to accepting the starting point as true.
5	Specified threshold for rejection of the starting point: the standard of proof.	Specified threshold for rejection of the starting point: the significance level
6	Recognition that one type of error, false conviction, is more regrettable than the alternative type of error (in criminal trials).	Recognition that one type of error, erroneous rejection of a true null hypothesis, is more regrettable than the alternative type of error.
7	Different thresholds reflect different disutilities for errors in different situations: beyond a reasonable doubt, clear and convicting, preponderance.	Different thresholds reflect different disutilities for errors in different situations: significance levels can be set at different values.

[1—Law] The decision choices in trials, both criminal and civil, can be represented by the familiar 2×2 matrix on the left side of Fig. 1. For convenience, let us consider a criminal trial. At the conclusion of the trial, the factfinder must choose a verdict of ‘guilty’ or ‘not guilty’ on each count charged (or on lesser included charges). In reality, the defendant is ‘guilty’ or ‘innocent’, but that reality can never truly be known to the factfinder. The factfinder can know, and may consider, only the case-specific evidence that has been presented during the trial and the inferences that can be drawn from that evidence. The law appreciates that although a defendant can, in some ultimate reality, be innocent, the factfinder can never declare a defendant to be innocent but is in a position only to find that the evidence presented did cross the requisite threshold to guilt (and the defendant is therefore ‘guilty’) or that the evidence did not cross the threshold (and the defendant is therefore ‘not guilty’).

[1—Science] The decision choices in conducting scientific research can be represented by the 2×2 matrix on the right side of Fig. 1. At the end of an experiment, the researcher must decide whether to ‘reject the hypothesis of no effect’ (the null hypothesis) or ‘not to reject that hypothesis’. For example, the research might be testing whether or not a drug causes a harmful side effect. In reality, the drug does or does not cause harm, but that reality cannot truly be known to the researcher. The researcher can know only the evidence that has been developed through the study and the inferences that can be drawn from that evidence. Science can conceive of an ultimate empirical generalization on the issue, but ultimate knowledge is beyond our reach. In any given study, researchers are in a position only to say whether or not the evidence crossed the threshold to allow one to say that an effect was found or that the evidence did not cross the threshold, and therefore the hypothesis of no effect cannot be rejected.

		Unknown State of Reality	
		Guilty	Innocent
D e c i s i o n	G	<i>Correct Conviction</i>	<i>Erroneous Conviction</i>
	NG	<i>Erroneous Acquittal</i>	<i>Correct Acquittal</i>

		Unknown State of Reality	
		H ₀ is False	H ₀ is True
D e c i s i o n	Reject H ₀	<i>Correct Rejection</i>	<i>Error Type I p < .05</i>
	Don't Reject H ₀	<i>Error Type II</i>	<i>Correct Non-rejection</i>

FIG. 1. Comparing legal and scientific decision-making under uncertainty.

[2—Law] There are two ways a legal factfinder can be correct: the defendant is guilty and the factfinder decides to convict or the defendant is innocent and the factfinder decides to acquit. Furthermore, there are two ways the factfinder can be incorrect: the defendant is guilty but the factfinder decides to acquit or the defendant is innocent but the factfinder decides to convict.

[2—Science] Similarly, there are two ways a researcher can be correct: the phenomenon exists and the researcher rejects the hypothesis of no effect or the phenomenon does not exist and the researcher does not reject the hypothesis of no effect. Furthermore, there are two ways the researcher can be incorrect: the phenomenon exists but the researcher decides that the null hypothesis cannot be rejected or the phenomenon does not exist but the researcher rejects the hypothesis of no effect.³

[3—Law] A decision tool is needed with which to begin a trial—before any evidence is presented and before the factfinders know any of the adjudicative facts of the case. Factfinders are screened for their lack of knowledge about the specifics of the case. Moreover, they are instructed to presume that the defendant is innocent. That is the starting point: the ‘presumption of innocence’. The factfinder’s task is then to evaluate the evidence that unfolds during the trial and to assess whether it moves their judgement of guilt so far from the starting point that a conviction is warranted.

[3—Science] Scientific studies have a similar starting point: ‘the null hypothesis’. The null hypothesis usually reflects a starting assumption about the relationship between the variables under scrutiny: no correlation between variables and no difference between experimental and control groups as a function of different experimental treatments. For example, in an experiment where the independent variable is a drug administered or not administered to two groups of people (the experimental and control groups) and the dependent variable is how many colds each group catches in a season, the null hypothesis would state the assumption that there will be no difference in the mean number of colds in the experimental and control groups. Expressed as an equation, that would be: H₀: M_E = M_C. That is an assumption, a temporary stance, not a finding, not a conclusion. If sufficient evidence accumulates in the study to reject the null hypothesis, the researchers can conclude

³ Scientists have the advantage of being able to conduct multiple studies and learning from the set of studies by performing either a traditional literature review or a meta-analysis. Though replication is less likely in the law, the law is not completely lacking in the ability to (more or less) replicate a trial—as when judges order new trials, which they do in some circumstances. Moreover, some scholars have put the theoretical possibility of numerous (or infinite) replications of legal trials to good use (e.g. Saks and Blanck, 1992). For present purposes, however, we focus on the situation where one study or one trial is being conducted at a time.

that the independent variable has an effect on the dependent variable. If the evidence is insufficient, then the null hypothesis is not rejected. Researchers might already know something about the drug, be that knowledge theoretical or empirical. But the analysis of study data does not incorporate that knowledge. It begins with an assumption of no difference, of no effect and of no relationship between independent and dependent variables.

[4—Law] The impropriety—the impossibility—of declaring a defendant ‘innocent’ deserves further comment. There are formal reasons for the inability to declare innocence. Factfinders begin the trial process with no information relevant to the disputed facts of the trial and are instructed to assume a starting point in which they presume innocence.⁴ The factfinder’s task is to evaluate the extent to which the trial evidence moves the decision away from that starting point and along the scale towards guilt. Even if by the close of evidence factfinders judge the evidence not to have moved off of the starting point at all, innocence is an assumption, a temporary stance and not a finding. If insufficient evidence of guilt is presented, the factfinders can declare the defendant to be ‘not guilty’. If the defence were to offer strong evidentiary support for innocence, the factfinder still is not permitted to declare the defendant innocent; that option does not formally exist. The most a factfinder can say is that the defendant was found ‘not guilty’.⁵ Practical, in contrast to formal, reasons also explain the inability to declare innocence. If evidence pointing strongly in the direction of innocence were uncovered (e.g. exculpatory DNA evidence or an airtight alibi), charges would rarely if ever be brought by the prosecution, would be dropped or would be dismissed by the court. So the factfinder would never be in a position to assess proof of innocence. The defence in a criminal trial is most often in the position of challenging the prosecution’s evidence of guilt, showing that it does not cross, or approach, the threshold for concluding that guilt has been established.

[4—Science] The formal impossibility of accepting the null hypothesis as true—of declaring ‘no difference’ or ‘no effect’—also deserves further comment. In a scientific study, a failure to reject the null hypothesis is understood to be quite different from accepting the null hypothesis as true. Non-differences can result for numerous reasons other than the reality that no actual relationship exists between two variables. These include independent variables that are insufficiently well manipulated or do not adequately operationalize the construct under study, dependent measures that do not capture the construct or are insufficiently sensitive, a study sample size that is too small, statistical tests that have too little power to detect a true effect, confounding that suppresses the relationship between the independent and dependent variables, and other problems.⁶ In short, both science and the law regard their respective decisional starting points as formally unable to serve as conclusions.

[5—Law] Another decision tool is needed: a threshold which, if crossed by the evidence, changes the verdict from the non-guilty presumption to a guilty verdict. For the law, that is the ‘standard of

⁴ Empirical research suggests that mock jurors accomplish this by according essentially zero weight to any beliefs they might be entertaining at the start of the case. The weights as well as the scale value of beliefs change as evidence unfolds. See *Ostrom et al. (1975)*.

⁵ The closest any Anglo-American trial system we are aware of comes to recognizing innocence is the Scottish trial which, in addition to guilty and not guilty, offers the option of ‘not proved’. ‘Not proved’ is equivalent to what in the English and American systems is termed ‘not guilty’. Thus, in the Scottish system, concluding that a defendant is ‘not guilty’ is tantamount to saying the defendant is innocent. See *Hope et al. (2008)*.

⁶ Though, as a formal matter, the null hypothesis can never be proven, as a practical matter, if enough well-designed and well-conducted studies of the same phenomenon are conducted, researchers can come to believe that X does not cause Y. Similarly, consider *Popper (1989)* (arguing that theories can never be proven true, but can only be disproved, though a theory that has survived many sound attempts at disproof comes to be regarded as valid).

proof'. In criminal cases, that standard is the 'threshold of reasonable doubt'. Factfinders are to deliver a verdict of 'guilty' if and only if the evidence and inferences from that evidence are sufficient to cross the threshold of reasonable doubt. In civil cases, of course, the threshold is set lower ('preponderance of the evidence' or 'the greater weight of the evidence'), where the presumption of non-liability favouring the defendant may be rejected on a mere tipping of the balance in favour of the plaintiff. In certain other kinds of cases, an intermediate threshold ('clear and convincing evidence') must be crossed in order to reject the presumption of non-liability.⁷

[5—Science] Science, similarly, defines a threshold beyond which the evidence changes the conclusion from non-rejection of the null hypothesis to rejection. For science using statistical hypothesis testing, that is the 'significance level' (also known as the alpha level or p level). Researchers are to conclude that a relationship between an independent and dependent variable exists if and only if the evidence crosses the line defined by that significance level. A typical significance level is $p < .05$, indicating that if the difference between the experimental and control groups are merely the result of random sampling error, a result as far from expected under the null hypothesis as the observed result would occur in fewer than 5 out of every 100 samples. Because scientists typically quantify their observations, they can use those measures to calculate the risk of a false rejection of the null hypothesis and that risk of error (in the present example, and usually) is kept below 5%.⁸

[6—Law] The law is not indifferent to the two kinds of errors described above. It recognizes that the two kinds of error stand in a ratio to one another. Erroneous convictions have a different value than erroneous acquittals. One might say they have disutilities of different magnitudes. Or, most simply, that one type of error is more serious, threatens valued interests more, than the other type of error. Most readers will be familiar with Blackstone's maxim that, 'It is better that ten guilty persons escape than that one innocent suffer' (Blackstone, 1765–1769, Book 4, chapter 27). Taken literally, this suggests that the law prefers 10 or more erroneous acquittals to 1 erroneous conviction, though perhaps Blackstone's comment is more figurative and represents only a general concept: that the law regards erroneous convictions as causing considerably more disutility than erroneous acquittals do.⁹ Other legal cultures, quite different from the Anglo-American system, have developed similar notions of the need to find the right ratio between erroneous convictions and erroneous acquittals.¹⁰

⁷ A number of efforts have been made to transform legally defined thresholds of proof into statements of probabilities. See e.g. Kagehiro (1990), Gastwirth (1992), Weinstein and Dewsbury (2006) and *United States v. Fatico* (1978) (in which Judge Weinstein reports the results of a poll of his judicial colleagues regarding the probabilities they see reflected in the conventional proof thresholds: preponderance, clear and convincing, beyond reasonable doubt).

⁸ Of course, features of significance tests are not everything. Sample size, magnitude of effect and error variance affect whether or not a study results in a significant difference. The larger a sample, the greater the effect size, and the smaller the error term, the more likely a statistically significant difference is to be found. In respect to these elements, only very rough analogies can be found in the legal trial: more witnesses, more compelling evidence and less contradictory evidence are more likely to lead to guilty verdicts. Moreover, these effects are psychological rather than statistical: responses of factfinders to information that they must integrate in order to reach an overall conclusion.

⁹ This also reveals that the law is not unaware of the risk—indeed, the inevitability—of error, in its procedures. See Risinger (2006–2007).

¹⁰ Other ratios have been suggested. Hale in 1736 suggested 5:1 ('... it is better five guilty persons should escape unpunished, than one innocent person should die'). Fortescue, writing in 1616, suggested 20:1 ('I would rather wish twenty euill [evil] doers to escape death through pittie, then one man to bee vniustly [unjustly] condemned'). Benjamin Franklin proposed 100:1. Maimonides, in the 12th century, argued for 1000:1. In Genesis, God agrees to spare the city of Sodom if 50 righteous can be found there, and Abraham persuades God to change that to 10 (either of them being a ratio of (perhaps) hundreds or even thousands to 1). The preceding examples are drawn from Volokh (1997), who provides quite a comprehensive review of court rulings, biblical references, historical views and other sources. Volokh found the ratio to range from 1:1 to 5000:1.

[7—Law] By changing the standard of proof, the law is changing the ratio of the two kinds of errors.¹¹

[6—Science] Similarly, science is not indifferent to the two kinds of errors discussed above. It recognizes that the two kinds of error stand in a ratio to one another. Erroneous rejections of the null hypothesis lead to conclusions that phenomena exist when they do not actually exist, and theories incorporating such erroneous findings would be constructed or renovated in error. That would lead to a great deal of error and confusion, so false rejection of the null is unwelcome to science, and therefore the risk of such errors is kept low (below 5%). Science tends to be conservative in the sense that it would prefer to overlook phenomena that actually do exist (non-rejection of a false null) than to recognize as true phenomena that do not actually exist (rejection of a true null). [7—Science] Scientists and statisticians realize that where they set the threshold for rejecting the null hypothesis changes the risk of one type of error relative to the other type of error.¹²

All told, the law's scheme for making difficult and consequential decisions in the face of uncertainty is no less sophisticated than the one used by scientists and statisticians. At bottom, the two are virtually twins.

3. Historical evolution of the two models

We could stop at this point and rest content to note these interesting similarities in two fields usually regarded as so different. But it is hard to resist the temptation to enquire into the origins of these two approaches to decision-making under uncertainty, which are really the same approach. Did law and science inform each other in any way or did their models evolve separately? Looking into the history of their origins provides insight into the answer.

3.1 *Origins of the presumption of innocence and the reasonable doubt standard of proof*

In law, the presumption of innocence has ancient origins. Langbein has summarized: the presumption of innocence 'was known from classical Roman Law and had been reinvigorated in the natural law literature of the Seventeenth Century. English juristic writers subscribed to it, from Fortescue to Coke to Blackstone' (Langbein, 2003, p. 262) (notes omitted).

The reasonable doubt standard of proof developed later. Although glimmers of the notion appear inconsistently in cases going back centuries,¹³ the rule did not crystallize in Anglo-American law

¹¹ In science, signal detection theory, a well-developed quantitative theory, describes these changes in ratios as a function of differing decision thresholds. See Phillips *et al.* (2001) and Wickens (2001).

¹² The description in this paragraph pertains most clearly to pure science, the enterprise of knowledge building for its own sake. In that endeavor, avoidance of false rejections of the null is understandably greatly favoured. But consider an instance of applied science, e.g. where one is testing a cure for a dreaded disease. In that context, it would make sense to run a greater risk of erroneously concluding that a treatment works when in reality it does not (and such interests would be effectuated by setting the p level at, say, $p < .10$ or $p < .20$, or balancing a type I error of .05 by a type II error that is not more than .10) than to erroneously conclude that a treatment does not work when in reality it is effective. Nevertheless, in practice, researchers of most kinds in most situations usually adhere to the conventional p levels of 5 or 1%. And, thus, the law has shown similar sophistication in its development of different decision thresholds that are routinely employed—preponderance, clear and convincing, beyond a reasonable doubt—for different categories of trials having different utilities and disutilities.

¹³ Jurists in criminal trials sometimes made such pronouncements as that the evidence must be 'clearer than the noonday sun' to convict a defendant (i.e. admitting of no, or virtually no, doubt). For example, Julius Clarus (1585) ('Debent autem esse in criminalibus probationes luce meridiana clariores... [e]t hoc omnes sciunt et dicunt') (quoted in Whitman, 2008). See also Morano (1975), finding reasonable doubt employed as early as 1770 in Massachusetts, in the Boston Massacre trials.

until the last quarter of the 18th century.¹⁴ The reasonable doubt standard appears to have become a familiar concept in English trials from the mid-1780s, though it took somewhat longer to become established as a rule of law.¹⁵

Whether the beyond-a-reasonable-doubt standard of proof in law emerged in the late 18th century or sooner than that, it clearly preceded the emergence of parallel concepts in science and statistics and therefore could not have derived from these non-legal uses. The social needs that gave rise to the reasonable doubt standard, as understood by recent historical scholarship, are worth examining.¹⁶

The reasonable doubt standard appears to have originated not to protect criminal defendants from erroneous conviction, but as one of the numerous responsibility-shifting and agency-denying devices adopted by the law to benefit the decision-makers, insulating them from moral responsibility so that they would be less fearful of delivering the decisions they were called upon to make. More specifically, the reasonable-doubt standard developed as a device for giving jurors what Whitman terms the 'moral comfort' to convict.

Until the early 13th century, judges in Christian Europe were fearful about performing their role because doing so violated New Testament injunctions against judging (Found in the 'Gospel of Matthew') and of causing bloodshed (Whitman, 2008, chapter 2). Christians of that time and place believed that taking any part in causing the death of an accused, or inflicting other blood punishments, however justifiable, might lead to eternal damnation. Bloodshed was viewed as a moral pollution that barred those responsible from receiving communion for 3 years and necessitated their purification. Even soldiers who killed in legitimate defence were considered to have become morally polluted.

For centuries, the solution for judges was the trial by ordeal. By subjecting defendants to ordeals—e.g. throwing them into water to see if they sank (indicating innocence) or floated (indicating guilt)—judges could avoid the responsibility and moral consequences of deciding cases by invoking God's judgement instead of their own. Such divine invocation, however, required the participation of the clergy. This passed the moral jeopardy from judges to clerics, who feared that if the procedure were not correct, and the ordeal led to a verdict that did not accurately reflect God's judgement, then it was the priests who would become tainted. The Church's fears about the moral pollution of its priests led the Fourth Lateran Council of the Christian Church to safeguard the purity of the clergy by adopting Canon 18, which prohibited priests from participating in judicial ordeals.

What were judges and courts to do now? New solutions were needed to elude the moral consequences of decision. In Continental Europe, Canon 18 led to the development of inquisitorial judicial procedures. By 'enquiring' of persons suspected of crimes, the process did not require the participation of accusers or witnesses. So long as the judge did not supply any of the evidence, or rely on any personal knowledge, judges were considered to be maintaining a safe moral distance from the bloody consequences of judicial decisions. The defendant's confession led to the verdict and caused the punishment. To maintain proper moral distance, judges needed to follow procedural rules rigidly, and canon lawyers developed the rule 'in dubio pro reo' (in doubt you must decide for the defendant) (Whitman, 2008, p. 122). This rule 'created a form of protection for the accused that grew out of the familiar fear that the judge might make himself into a murderer' (Whitman, 2008, p. 133).

¹⁴ Langbein (2003) states: 'In routine criminal adjudication at the Old Bailey, we find scant indication of any standard of proof being the subject of jury instruction until the last quarter of the eighteenth century' (p. 262).

¹⁵ That is, that it moved from application according to the discretion of individual judges to being invariably applied by all judges.

¹⁶ See Whitman (2008) and Gallanis (2009). These are our sources for the discussion that follows.

In England, the extinguishing of the institution of the ordeal led to the invention of the jury—first consisting of men from the community with knowledge of the facts at issue (self-informing juries), and eventually¹⁷ to juries constituted and operated in ways more familiar to us today. Passing the moral risks of decision-making to jurors solved the judge's problem by creating problems for the jurors. Various rules came into being to help ensure that jurors would, when appropriate, convict defendants of serious crimes. These included fines and imprisonment for jurors who refused to render a verdict, special verdicts (facts found but no judgements made), immunity from penalties for erroneous convictions, allowance of verdicts other than blood punishments and the development of new punishments that avoided bloodshed. From 1718 until the outbreak of the American Revolution, transportation to the American colonies was the most common punishment imposed by English courts.

By rendering transportation impossible, the American Revolution created a serious problem for English courts, whose juries again faced the moral risks of being asked to reach verdicts that would lead to blood punishments. The solution was the development of the reasonable doubt standard of proof, 'a formula intended to ease the fears of jurors who might otherwise refuse to pronounce the defendant guilty' (Whitman, 2008, p. 193). If proper procedures were followed, and if there were no reasonable doubt as to the accused's guilt, then jurors could be comforted that the defendant was convicted by the law and the facts and not by the jurors themselves. Thus, although today the law views the standard of proof as a device to protect defendants from wrongful conviction, the historical development suggests that its real purpose was to protect jurors from other-worldly consequences of their decisions.

Though the law's version of these concepts was in place by the end of the 18th century, science and statistics had to wait until the curtain rose on the 20th century.

3.2 *Origins of statistical significance testing and the null hypothesis*

Statistical analyses, largely of descriptive kinds, have been employed by researchers in science since at least the middle of the 18th century (Stigler, 1992). Among the first scientists to employ statistics were astronomers, who used them to determine the coefficients in the equations for the movement of planets. In 1835, the French physician Pierre Louis used statistics to show that bleeding patients was not as effective a way to treat pneumonia as was then believed and, in 1855, the English physician John Snow analysed a natural experiment to demonstrate that cholera likely is an infectious disease, rather than a result of the body's unbalanced humours (Freedman, 1999). In 1860, the psychophysicist Gustav Fechner introduced statistics to the field of psychology in his experiments on sensation and perceptual thresholds of weight detection (Stigler, 1992). Despite a relatively long history in the sciences, however, statistical analyses did not much resemble the current model for hypothesis testing until the early 20th century.

In 1908, a groundbreaking statistical paper emerged from an unlikely enterprise (Box, 1987). In breweries all across Europe, despite precise techniques and lengthy apprenticeships, beer-making was still a hit-or-miss process. At Ireland's Guinness Brewery, however, Cecil Guinness was becoming the first beer-maker to appreciate the value of scientific experimentation in the pursuit of a consistent and superior brew. Guinness implemented the forward-thinking policy of finding the best and brightest chemistry graduates from Oxford and Cambridge, and hiring them as brewers. In 1899, William Sealy Gosset, a recent Oxford graduate, was brought on as one of these recruits.

¹⁷ 'Eventually' being the 19th century.

Gosset and his colleagues began experimenting on the effects of a variety of factors—hops, barleycorn size, amount of rainfall, temperature and so on—on the quality of the beer and soon realized that they could not use their results to make conclusive statements. The problem was that their experiments did not have enough data points to use the standard statistical analyses of the time. Statisticians in the early 1900s used formulas that assumed the standard deviation of the sample was the same as the true standard deviation of the population, which was not problematic when the sample sizes were large enough. But Gosset realized that this assumption was not valid for his small experimental sample sizes. After consultation with Karl Pearson, a leading biometrician, Gosset developed a statistical analysis that took into account the size of the samples and allowed valid comparison of two smaller groups. Guinness did not want Gosset to publish the paper under his own name, apparently fearing that doing so would reveal that the secret behind Guinness's famous brew was scientific testing and statistics. Therefore, Gosset had the paper published under the name 'Student', and his influential and still widely used t -test entered the world as 'Student's t -test'.

In 1912, while Gosset continued his work in brewing and statistics, a brilliant Cambridge mathematics undergraduate named Ronald A. Fisher was developing his own statistical formula. This formula, which later became known as the method of maximum likelihood, used a different calculation for standard deviation than the one Gosset had used. Fisher contacted Gosset to discuss the discrepancy and used mathematical proofs to show that his own formula was the correct one. This started Fisher down a highly prolific road of revolutionary mathematical research and proofs, and in 1925, he published his notable book, *Statistical Methods for Research Workers*. The book was intended, as the name implies, for researchers; so, while the formulas throughout the book were highly complex, readers did not need to understand the mathematics involved to learn how to use the tools they provided.

In his book, Fisher extended application of Student's t -test to testing regression coefficients and analysis of variance, and he also advocated a standard criterion for determining whether a test was significant or not (Lehmann, 1993). When testing a null hypothesis, researchers run statistical analyses that yield a p value, which is the probability that the data they gathered would have been found in a population in which the null hypothesis was true.

Previously, researchers would look at their p values and determine, somewhat arbitrarily, whether they were 'small enough'. Fisher recommended using a standard 0.05 cut-off point to be known as the 'alpha level'; a researcher could set his alpha level at 0.01, instead, if he wanted his test to be very stringent. Fisher's book was a great success, and the use of his statistical formulas became commonplace in many areas of research, particularly the social, behavioural and biological sciences.

Not long after the publication of Fisher's book, the statisticians Jerzy Neyman and Egon Pearson (son of the aforementioned biometrician Karl Pearson) wondered if there were ways to build on Fisher's methods. They were impressed with Fisher's idea of the standard p value cut-offs, and furthermore, they realized that there was a second side to the usual statistical story of accepting or rejecting a single hypothesis: if a hypothesis is being tested, and the decision is made to reject it, then theoretically an alternative hypothesis is being supported. These competing hypotheses became known as the 'null hypothesis', the hypothesis of no difference and the 'alternative hypothesis', which states that the two groups are different from each other. Although the alternative hypothesis is the hypothesis of interest, the null hypothesis is the one that is tested. Thus, hypothesis testing is indirect.

In two influential papers, [Neyman and Pearson \(1928, 1933\)](#) extended this line of reasoning, arguing that in hypothesis testing, two types of error are possible. The first, which they named ‘Type I error’, is rejection of the null hypothesis when the two groups are in fact the same. This is also known as a ‘false positive’ because it leads researchers to conclude that their alternative hypothesis is supported when it is not really true. The second, ‘Type II error’, is failure to reject the null hypothesis when the two groups are in fact different from each other. This is known also as a ‘false negative’ because a researcher making this error would reject the alternative hypothesis though it actually is true.

[Neyman and Pearson \(1928, p. 291\)](#) acknowledged that statistics would not allow researchers to ascertain the answer as it truly is in the population, but, ‘without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not too often be wrong’. They argued that the best hypothesis test would be one that balanced the two types of errors at an acceptable level, and they concluded that Fisher’s alpha level of 0.05 was appropriate. This general framework for hypothesis testing is remarkably similar to the one that had been in place in the legal world for more than a century, and it remains the predominant type of hypothesis testing used today.

Of interest to the present paper is whether Neyman and Pearson or any of the other scientists or statisticians working on this model of hypothesis testing were aware that a very similar model already existed in law. Some statistical papers made reference to the law as an illustration of decision-making under uncertain circumstances. [Neyman and Pearson \(1928, p. 296\)](#), e.g. refer to the mathematician Laplace’s question regarding judge and jury decision-making and briefly discuss the implications of errors of sending innocent people to prison versus setting guilty people free. However, we could find no recognition of the precise and nuanced legal conceptions that might have led to the similar conceptions adopted by the statisticians. The statistical literature we reviewed contained no reference to legal arguments, no citation of judicial opinions or legal commentary or legal professionals and no use of legal terminology. This suggests that the statisticians had only a lay understanding of the model employed in legal proceedings and were unaware of the strong similarity between the model being developed in statistics and the one already in place in law.

4. Conclusions

Our search for indications of any shared ancestry in the development of the legal and the scientific models of decision-making under uncertainty turned up no evidence of cross-fertilization. We found no indications that members of either field even considered the parallel problems each faced in their decision-making models at the time of their conception nor is there much indication that even today more than a handful of members of either field are aware of the striking similarity in their solutions. If scientific decision-making was not informed by the legal decision-making model that had become crystallized more than a century earlier, then two remarkably similar models arose independently in wholly unrelated fields. This in itself is noteworthy and represents an unappreciated commonality between law and science.

Moreover, such similar developments are not inevitable any time, in any field, whenever decision-making under uncertainty must be confronted. After all, the details of the trial process we describe crystallized relatively recently in the history of the Anglo-American trial, not becoming established in their recognizable form until after the 1780s. Before that, other solutions dominated. For example,

at an earlier stage of trial evolution, jurors were self-informing: witnesses and factfinders rolled into one. Such an arrangement precludes the kind of control over decisions achieved by the more modern trial system which has been the focus of this paper.¹⁸

4.1 *Biological evolution and cultural evolution*

We might analogize what has happened to something that occurs in biological evolution. The development of similar adaptations in two unrelated species is termed ‘convergent evolution’.¹⁹ Such evolution occurs when these unrelated species experience the same kinds of selection pressures, which are then met by the each species in similar ways. A familiar example is convergent evolution in birds and bats. Both have wings, even though their most recent common ancestors did not, indicating that both species separately responded to their environments by developing this adaptation. The specific characteristics of their wings differ—for example bird wings have feathers and bat wings have fur—but the function of wings is the same for both species. The biological world presents many instances of such similarity of response to similarity of environment.

Perhaps this concept from biological evolution can be borrowed to characterize something that happens in the ‘evolution’ of intellectual inventions and suggest that where the same underlying problem is encountered in different subject matter areas, similar solutions tend to be found. Though the solutions might appear different on the surface, a deeper look can discover that their structures as well as their functions are essentially the same.

4.2 *Differences amid the similarities*

Nothing we have written should be taken to suggest that the two systems do not also possess differences or that they are not employed differently or that they will not evolve beyond their present forms in different directions. After all, bats and birds, aloes and agaves and so on are not identical to each other.

Scientists follow the general procedures of hypothesis testing quite rigorously. That is, they perform significance tests carefully and regularly in their research. The comparable decision-making system in legal trials is more of a heuristic, a general model that is used by judges and imposed on jurors, with limited elaboration through instructions, with the details of actually carrying out the ‘test’ left largely unspecified.

Relatedly, while scientific studies often involve true experiments or systematic sampling, trials are more akin to observational studies—more like epidemiological studies than clinical trials or random sample surveys. This difference underscores the substantial imperfections in the data (evidence) gathered for trials and associated risks of both measurement error and errors of inference.

Statisticians have a large collection of different statistical significance tests adapted to different kinds of data and employed in specified circumstances. Judges and jurors, by contrast, do whatever

¹⁸ See discussion of this history in Shapiro (1993).

¹⁹ Both parallel and convergent evolution refer to the result of evolutionary processes that brought unrelated species to have similar features. If their respective ancestors had similar features which evolved separately so that their descendants had new different similar features, the process is termed parallel evolution (same evolves into same). But if their respective ancestors had different features which evolved separately so that their descendants came to have new features that were similar, the process is termed convergent evolution (different evolves into same). See Futuyama (1997). Convergent evolution seems a better fit to the analogous process that has occurred in the cultural evolution of law and science, considering that the purposes of the two started out being so different, before eventually arriving at the same structure for the same purpose.

they will or can with the varied kinds of evidence marshaled by the parties, while remaining within the general framework that has been the subject of this article.

The kinds of evidence used by statisticians are overwhelmingly quantitative—measures or counts of similar things, usually sampled in systematic, representative fashion from the population of similar things being studied. By contrast, the kinds of evidence brought to court are of a varied and messy lot: witnesses, documents, exhibits, expert opinions and more.

Furthermore, the evidence gathered by lawyers is anything but representative of the complete ‘population’ of evidence relevant to a case. The adversary process expects, almost demands, that the parties seek out the most favourable evidence they can find supporting their respective positions. Thus, while scientists usually bring to statistical hypothesis testing data which form normal balanced distributions, which in experiments have been randomly assigned to different experimental conditions; to the trial process lawyers usually offer up bimodal distributions, with each side presenting evidence biased toward its respective position.

Despite such differences in some of the details, the important similarities remain strong.

4.3 *New approaches to decision-making under uncertainty?*

Some might argue that the worlds of legal trials and of scientific hypothesis testing have moved beyond the tools we have described in this paper. They might point in particular to applications of Bayes’ Theorem in both worlds, seen by Bayes’ proponents as far more nuanced and adaptable and (in some sense) accurate.²⁰

But as illuminating as Bayes can be, the role it has come to play in the testing of statistical hypotheses in scientific research and in legal trials remains limited. Bayesian approaches to analysing data are few and far between in reports of scientific research, while hypothesis testing following the general model we have described remains the dominant tool.²¹ Similarly, in legal trials, while such suggestions have been discussed by scholars for decades, and Bayesianism has been one of the centerpieces of ‘the new evidence scholarship’,²² few if any practical applications of Bayesian approaches have come to displace the conventional understanding of how the law conceives of its factfinders reaching their verdicts.

The work that factfinders do in the course of reaching their verdicts (within that larger traditional decision-making framework) could, it has been argued by numerous scholars, be described and guided by Bayesian analysis.²³ But psychological research on jury decision-making has found this approach to be a poor description of what jurors actually do and cannot predict the results they reach (Sanders *et al.*, 2009). Nor is it reflected in anything that jurors are expected to do or are advised to do by courts.²⁴

²⁰ See discussion and debate in *Symposium* (1986), *Boston University Law Review* and *Symposium* (1991), *Cardozo Law Review*.

²¹ The best support for the textual statement would be found by looking through current scientific journals or the syllabi of basic statistics courses or statistics textbooks. There one will find massive attention to conventional hypothesis testing and little if any use of, or discussion of, Bayesian alternatives.

²² See symposia cited in note 20 and discussion in *Park and Saks* (2006).

²³ That is, factfinders each begin at some starting point (presumably something close to the presumption of innocence) and update their estimates of the probability of guilt as each new piece of evidence is provided. (Or a similar process in the civil context.)

²⁴ See also *R v. Adams* (1996), in which an English court explicitly rejected Bayesian thinking offered to a jury. We must acknowledge, however, the value of much Bayesian thinking and point to the role that it can usefully play in the law. As one

In short, neither in law nor in science has Bayesian thinking replaced traditional decision-theoretic concepts for making decisions under uncertainty. Had that occurred, we would be pointing out either that this example of cultural evolution in law and science has continued further along parallel lines or that it has diverged, as the case might be.

4.4 *Biological evolution versus cultural evolution*

Analogies have their limits. Clearly, the developers of these legal and scientific models had advantages over biological evolution. Biological evolution depends on random mutation and is a relatively slow process, whereas the development of trial procedures in law was a matter of social problem-solving embedded in a larger culture and the developments in statistical theory by Fisher, Neyman, Pearson and others involved forethought, design and revision.

Though not the products of randomness-and-selection, human rationality and creativity are inevitably bounded, and so the respective models might be flawed. The fact that the models are so similar does not speak to their inherent suitability or adaptiveness. That both have survived for long periods of time in their respective fields does, however, suggest that they function well to meet the problems they were developed to solve and continue to do so. But in the future, they might evolve further—and might diverge from each other—in order to further improve their function in their respective domains.

But, for the present, the striking similarity of law and science in their respective models of decision-making under uncertainty is worth marvelling at.

Acknowledgement

The authors thank Jonathan Rose for his very helpful guidance concerning relevant sources of legal history, as well as several anonymous reviewers and the editor for numerous helpful suggestions.

REFERENCES

- ARKES, H.R., & MELLERS, B.A. (2002) Do juries meet our expectations? *Law and Human Behavior*, **26**, 625–639.
- BLACKSTONE, W. (1765–1769) *Commentaries on the Laws of England*. Clarendon Press.
- BOX, J.F. (1987) Guinness, Gosset, Fisher, and small samples. *Statistical Science*, **2**, 45–52.
- BROWN, L.D. (2000) An essay on statistical decision theory. *Journal of the American Statistical Association*, **95**, 1277–1281.
- COHEN, J., COHEN, P., WEST, S.G., & AIKEN, L.S. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

of the authors has noted elsewhere, the best uses of Bayesianism have arisen outside of trial decision-making, in analysing the benefits and costs of existing and proposed rules. See Park and Saks (2006), pp. 991–992. Or for conceptualizing rules. See e.g. Lempert (1977). Moreover, some fields of expertise use Bayes' Theorem to analyse the evidence to which their expertise is directed, and the Bayesian analysis is part of the foundation for their opinions, such as in analysis of the meaning of DNA testing in parentage disputes. For example, *State v. Skipper* (1994) (in which the court wrote: 'The utilization of Bayes' theorem by the prosecution... permitted the introduction of evidence predicated on an assumption that there was a fifty-fifty chance that sexual intercourse had occurred in order to prove that sexual intercourse had in fact occurred... The fifty-fifty assumption that sexual intercourse had occurred was not predicated on the evidence in the case but was simply an assumption made by the expert'). See Morris and Gjertson (2009).

- CONNOLLY, T. (1987) Decision theory, reasonable doubt, and the utility of erroneous acquittals. *Law and Human Behavior*, **11**, 101–112.
- COWAN, T.A. (1963) Decision theory in law, science, and technology, *Science*, **140**, 1065–1075.
- COX, D.R., & HINKLEY, D.V. (1974) *Theoretical Statistics*. Chapman and Hall Ltd., London.
- FAIGMAN, D.L. (1999) *Legal Alchemy*. W. H. Freeman, New York.
- FREEDMAN, D. (1999) From association to causation: some remarks on the history of statistics. *Statistical Science*, **14**, 243–258.
- FUTUYMA, D.J. (1997) *Evolutionary Biology*, 3rd ed. Sinauer, Sunderland, Massachusetts.
- GALLANIS, T.P. (2009) Reasonable doubt and the history of the criminal trial. *University of Chicago Law Review*, **76**, 941–964.
- GASTWIRTH, J. (1992) Statistical reasoning in the legal setting. *American Statistician*, **46**, 55–69.
- GOLDBERG, S. (1996) *Culture Clash: Law and Science in America*. New York Univ. Press, New York.
- HAACK, S. (2009) Irreconcilable differences? The troubled marriage of science and law. *Law and Contemporary Problems*, **72**, 1–24 (Winter).
- HOPE, L., GREEN, E., MEMON, A., GAVISK, M., & HOUSTON, K. (2008) A third verdict option: exploring the impact of the not proven verdict on mock juror decision making. *Law and Human Behavior*, **32**, 241–252.
- KAGEHIRO, D.K. (1990) Defining the standard of proof in jury instructions. *Psychological Science*, **1**, 194–200.
- KAPLAN, J. (1968) Decision theory and the factfinding process. *Stanford Law Review*, **20**, 1065–1092.
- KEPPEL, G., & ZEDECK, S. (1989) *Data Analysis for Research Designs: Analysis of Variance and Multiple Regression/Correlation Approaches*. W.H. Freeman and Company, New York.
- LANGBEIN, J.H. (2003) *The Origins of Adversary Criminal Trials*. Oxford Univ. Press, New York.
- LEHMANN, E.L. (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association*, **88**, 1242–1249.
- LEMPERT, R.O. (1977) Modeling relevance. *Michigan Law Review*, **75**, 1021–1057.
- MORANO, A.A. (1975) A reexamination of the development of the reasonable doubt rule. *Boston University Law Review*, **55**, 507–529.
- MORRIS, J.W., & GJERTSON, D.W. (2009) Parentage testing. In Faigman, D. et al., eds., *Modern Scientific Evidence*. West, St. Paul, Minnesota.
- NEYMAN, J., & PEARSON, E.S. (1928) On the use the interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20A**, 263–294.
- NEYMAN, J., & PEARSON, E.S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character*, **231**, 289–337.
- OSTROM, T.M., SAKS, M.J., & WERNER, C.M. (1975) The presumption of innocence and the American juror. *Journal of Contemporary Law*, **2**, 46–54.
- PARK, R., & SAKS, M.J. (2006) Evidence scholarship reconsidered: results of the interdisciplinary turn. *Boston College Law Review*, **47**, 949–1032.
- PHILLIPS, V., SAKS, M.J., & PETERSON, J. (2001) Signal detection theory and decision-making in forensic science. *Journal of Forensic Sciences*, **46**, 294–308.
- POPPER, K. (1989) *Conjectures and Refutations: The Growth of Scientific Knowledge*, 5th ed. Harper & Row, New York.
- R V. ADAMS (1996) 2 Cr App R 467 (England).
- RISINGER, D.M. (2006–2007) Innocents convicted: an empirically justified factual wrongful conviction rate. *Journal of Criminal Law and Criminology*, **97**, 761–806.

- SAKS, M.J., & BLANCK, P.D. (1992) Justice improved: the unrecognized benefits of aggregation and sampling in the trial of mass torts, *Stanford Law Review*, 44, 815–851.
- SANDERS, J., SAKS, M.J., & SCHWEITZER, N.J. (2009) Trial factfinders and expert evidence. In Faigman, D. et al., eds., *Modern Scientific Evidence*. West, St. Paul, Minnesota.
- SHAPIRO, B.J. (1993) *Beyond Reasonable Doubt and Probable Cause: Historical Perspectives on the Anglo-American Law of Evidence*. Univ. California Press, Berkeley, California.
- STATE V. SKIPPER (1994) 228 Conn. 610, 637 A.2d 1101 (1994).
- STIGLER, S.M. (1992) A historical view of statistical concepts in psychology and educational research. *American Journal of Education*, 101, 60–70.
- SYMPOSIUM (1986) Probability and inference in the law of evidence: the uses and limits of Bayesianism. *Boston University Law Review*, 66, 377–952.
- SYMPOSIUM (1991) Decision and inference in litigation. *Cardozo Law Review*, 66, 253–1079.
- THOMAS, E.A.C., & HOGUE, A. (1976) Apparent weight of evidence, decision criteria, and confidence ratings in juror decision making. *Psychological Review*, 83, 442–465.
- TRAFIMOW, D. (2003) Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayes's theorem. *Psychological Review*, 110, 526–535.
- UNITED STATES V. FATICO (1978) 458 F. Supp. 388 (1978).
- VOLOKH, A. (1997) n guilty men. *University of Pennsylvania Law Review*, 146, 173–216.
- WEINSTEIN, J.B., & DEWSBURY, I. (2006) Comment on the meaning of proof beyond a reasonable doubt. *Law, Probability and Risk*, 6, 167–173.
- WHITMAN, J.Q. (2008) *The Origins of Reasonable Doubt: Theological Roots of the Criminal Trial*. Yale Univ. Press, New Haven, Connecticut.
- WICKENS, T.D. (2001) *Elementary Signal Detection Theory*. Oxford Univ. Press, New York.