3-2-2017

Probability, Explanation, and Inference: A Reply

Michael S. Pardo
University of Alabama - School of Law, mpardo@law.ua.edu

Ronald J. Allen
Northwestern University - Pritzker School of Law, rjallen@northwestern.edu

Probability, Explanation, and Inference: A Reply

Ronald J. Allen
Michael S. Pardo

11 INTERNATIONAL JOURNAL OF EVIDENCE AND PROOF 307
(2007)

Probability, explanation and inference: a reply

By Ronald J. Allen and*
Wigmore Professor of Law, Northwestern University; Fellow, Procedural Law Research Center, China Political Science and Law University

Michael S. Pardo†
Assistant Professor, University of Alabama School of Law

The inferences drawn from legal evidence may be understood in both probabilistic and explanatory terms. Consider evidence that a criminal defendant confessed while in police custody. To evaluate the strength of this evidence in supporting the conclusion that the defendant is guilty, one could try to assess the probability that guilty and innocent persons confess while in police custody. Or one could make the same assessment based on any number of other characteristics shared by the defendant or the context of the confession. The problem of reference classes would arise quite readily because each of these different classes would likely yield different results, some of which will take one closer to, and some further away from, the correct conclusion of whether this defendant in fact is guilty. Alternatively, one could evaluate how well the conclusion that the defendant is guilty explains the evidence of the confession.[1] How well the defendant's guilt explains the evidence will depend on the strength of alternative explanations such as whether a false confession was coerced, or the defendant was trying to protect another person from conviction, or the police are lying about whether a confession was ever given, etc.

* Email: rjallen@law.northwestern.edu. Professor Allen is indebted to the Julius Rosenthal Foundation and the Searle Center for support of the preparation of this article.
† Email: mpardo@law.ua.edu. My thanks to Dean Ken Randall and the University of Alabama Law School Foundation for generous research support.

1 Both approaches involve inductive inferences; the probability approach employs enumerative inferences and the explanation approach employs abductive inferences.

To be sure, evidence expressed in probabilistic terms may sometimes be helpful in evaluating the plausibility of completing explanations. We nonetheless contend that such explanatory considerations better explain important aspects of the nature of juridical proof, including the value of legal evidence and standards of proof, than competing probabilistic accounts[2] Probability approaches to judicial evidence and proof, including formal mathematical models, are useful to the extent that they contribute to these explanatory considerations and less so when they do not. We elaborate on these points in reply to the four papers in this symposium. The responses by Nance, Colyvan and Regan, and Rhee, which we discuss first, focus primarily on the probability perspective. Laudan focuses on the explanatory perspective, to which we turn in conclusion.

Nance

Nance argues that the reference class problem does not preclude the use of formal models as tools to evaluate inferences. He suggests that we are unnecessarily sceptical in rejecting the formal analysis and that those modelling such evidence are aware of several of the limitations that we pointed out, including that: the correct reference class cannot be deduced from the environment; class selection involves judgment informed by inferential interests and epistemic values; probative value cannot be equated with random or arbitrarily chosen classes; and the need to consider multiple classes. Significantly, Nance suggests, in such models there is no 'claim or implicit assumption that a unique reference class is necessarily privileged'.[3]

First, we are not uncompromising sceptics, and we even concede that the formal models may be useful in evaluating evidence and that it may not be unreasonable for parties to argue, or fact-finders to evaluate evidence, along the lines suggested by such models. To say that such models may be useful is not, however, to accept them as the sole or even a particularly reliable method of discovering the truth. Our objection is to scholarship arguing just that such models establish the correct or accurate probative value of evidence, and thus implying that any deviations from such models lead to inaccurate or irrational outcomes. We think it telling that Nance labels critics of this one methodology as sceptics, as though there were no other paths to truth and no other tools of rationality. This is a very narrow view of rationality and inquiry, and we do indeed reject it. Formal methods of the sort Nance is espousing are powerful but within limited

2 We discuss this conclusion and these issues in more detail in Michael S. Pardo and Ronald J. Allen, 'Juridical Proof and the Best Explanation' *Law and Philosophy* (forthcoming).

3 Nance, 'The Reference Class Problem and Mathematical Models of Inference', above at 259.

domains, and most of human experience, and thus most inferential problems, do not lie within those domains.

Secondly, Nance's protestations to the contrary, we think it obvious that the models in the articles that we addressed do privilege certain reference classes, when such classes and such models are by no means the only reasonable way to evaluate the evidence. Thus, the strong conclusions drawn from the models are unwarranted. To demonstrate we focus again on the Nance and Morris DNA studies and the arguments by Finkelstein and Levin.

Nance responds that the 'main conclusions in the Nance and Morris studies' are 'that jurors tend to undervalue DNA match evidence relative to the stated Bayesian norms, but that the degree of undervaluation can be reduced by certain kinds of Bayesian instruction'.[4] The authors obviously think that reducing this 'undervaluation' is a good thing; they write that 'a careful use of such Bayesian methods in the courtroom can indeed assist the jury in reaching more accurate verdicts'.[5] Although Nance asserts that 'no fair reading' of the studies takes them as 'privileging' reference classes, nor as declaring juror conclusions to be incorrect or irrational,[6] how can one speak of *under*valuing the evidence or reaching *more accurate* verdicts without assuming that the models are providing a correct evaluation of the evidence? To assert an error in the experimental responses, the models must privilege some reference class to calculate the supposed correct response, whether the random-match and lab-error numbers given in the testimony (the first articulation) or those chosen by the witnesses (the second articulation). Our original discussion focused on the first articulation because it presented the reference class problem most vividly. For example, why should national statistics be accepted as a basis for calculating the probability that *this* lab at *this* time erred? As Nance points out in response, and as we discussed in our original article, one possible reason the subjects did not accept the testimony rates at face value may have been (surely intuitive) doubts about the choice of reference class. It strikes us that further study of why the jurors did what they did would be helpful.[7] But, even given that they did so, why in the second articulation should even their chosen numbers be taken at face value for modelling the 'correct' value? Why should

4 Ibid. at 269.

5 Dale A. Nance and Scott B. Morris, 'An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability' (2002) 42 *Jurimetrics Journal* 403, 445.

6 Nance, 'The Reference Class Problem and Mathematical Models of Inference', above at 269.

7 For a discussion of recent empirical literature on the different ways individuals interpret frequency data, see Aron K. Barbey and Steven A. Sloman, 'Base-rate Respect: From Ecological Rationality to Dual Processes' *Behavioral and Brain Sciences* (forthcoming 2007).

jurors not, for example, harbour second-order doubts about the previous numbers (again, which privilege a reference class), which they may have pulled out of thin air, by further discounting when assessing the evidence as a whole? Again, in our view, further research should attempt to understand how humans reason but, at least for the time being, has no warrant for directing them to reason in the particular ways endorsed by the researchers.

All the Nance and Morris study demonstrated was that the experimental subjects did not interpret and manipulate the information as the researchers would have liked them to have done. But the experiments did not purport to be about the subjects' ability to divine how the experimenters wanted them to act; the experiments purported to be about assessments of guilt and innocence. The experimenters did not tell the subjects to put aside their common sense and background knowledge and look only at, and try to manipulate in a formal fashion, the information they had in front of them; rather, the subjects were asked to provide an evaluation of the evidence. It is perfectly reasonable in doing so to treat this as a real problem of human inference involving all the complexity and concerns that any real inference under uncertainty about human affairs entails, and for the resolution of which formal modelling is of extremely limited utility. Of course, it is also possible that the experimental subjects acted quite irrationally, but this experiment does not establish or even provide good evidence of that. At most it points to further work to be done to uncover how the subjects reasoned.[8]

Similar conclusions apply to Finkelstein and Levin's example in which a witness testifies that the frequency of a carpet fibre found at the crime scene and matched to a carpet owned by the defendant is 1 in 500. Finkelstein and Levin present their example in terms of what a prosecutor may or may not reasonably argue, but they evaluate this based on what they believe the analysis has established as the correct probative value of the evidence. Their analysis is clear that, once the 1 in 500 figure is given, the calculated ratios (depending on whether the tests are all from the defendant's carpets or he is one of several suspects) apply to correct values of the

8 Nance asserts that we incorrectly accused the studies of suggesting juror irrationality. But he also describes the studies as exercises in the 'rationality constraints' imposed by probability theory. If subject answers are considered to be incorrect or undervalued to the extent that they deviate from the model's conclusions, and if complying with the constraints would lead to more accurate verdicts, we see in this the implications that subjects were not acting as rationally as they should be and that their rationality could be improved by the Bayesian instructions. This is all we meant by 'juror irrationality'.

evidence.[9] Yet the reference class problem makes such bold conclusions unwarranted. We have no idea about the class from which these statistics were taken (town, state, country, store, etc.). At one point, the authors refer to distribution in the town.[10] But why assume these statistics would apply randomly and uniformly to this town and within this town? Further investigation might reveal that nearly every house in this town has the carpet fibre. We are not suggesting that witnesses and parties may not explore these statistical and related issues. We are taking issue with the idea that the models have established the correct value of the evidence. To do so, however, the model must make several background assumptions regarding, for example, uniformity, randomness and independence of the evidence.[11] Different assumptions and therefore different models could be constructed, employing the same and different reference classes,[12] and each would yield a different answer to the value of the evidence.

Consider, for example, Finkelstein and Levin's assertion that, in the case where the prosecution tested 20 different houses, the results of the other 19 searches are irrelevant because, as they say, 'the question is not whether *any* one of 20 matches would be made, but whether Jones's fiber would match'.[13] Suppose, however, that there was a match in all 20 of the samples taken. Would it be 'rational' to use the 1 in 500 number as the 'correct' reference class to draw inferences? Almost certainly not. The results in our modified hypothetical indicate that the national figures relied on by the expert are uninformative about this town, or this neighbourhood, or this street, or this block, or whatever, pertinent to this case. This, of course, is our point. Finkelstein and Levin, like Nance, implicitly are asserting that there is a correct value of the evidence (one that would lead to more accurate results), which we dispute. They have merely latched on to one of many possibilities, and assumed

9 See e.g. Michael O. Finkelstein and Bruce Levin, 'On the Probative Value of Evidence from a Screening Search' (2003) 43 *Jurimetrics Journal* 265, 268: 'The hypothetical illustrates that the probative force of an evidentiary trace represented by the factor of 500 comes from narrowing the suspects to a small group out of a larger population' and 'Given the match, the odds that the fiber came from Jones's home (versus a different home) are 500 times greater than the odds would have been ignoring the match evidence. This is undoubtedly powerful evidence.' We read these conclusions to refer to the actual or correct value of the evidence, not just what it would be reasonable to argue. We think, by contrast, that it might be reasonable to doubt the power of the evidence.

10 Ibid. at 267, n. 9.

11 Ibid. at 268 ('The presence of other evidence does not change the analysis given above because the *increase* in probability that Jones's house was the source of the fiber associated with finding a match is not affected by the other evidence.').

12 For an illustration of how different models and auxiliary assumptions would lead to different conclusions in an actual case, see Ronald Meester, Marieke Collins, Richard Gill and Michiel van Lambalgen, 'On the (Ab)use of Statistics in the Legal Case Against Nurse Lucia de B.' *Law, Probability and Risk* (forthcoming 2007).

13 Finkelstein and Levin, above n. 9 at 267.

it to be uniquely correct without adequate justification, epistemic or otherwise. Nance says that we misunderstand the point of their paper, which is simply to show what to do on the assumption that their favourite reference class is correct. Our critique, by contrast, was trying to show that their demonstration is really quite beside the point. There will virtually never be a correct reference class to allow strong statements about how the evidence should be handled. Thus, rather than our missing the point of this paper, Nance missed the point of our critique. The lesson to be learned from studies like Finkelstein and Levin's is not about misevaluating evidence, but is instead a cautionary tale about inadvertently misevaluating the evidentiary process.

Colyvan and Regan

As we stated in our original article, we agree with the main conclusion of Colyvan and Regan that *United States* v *Shonubi*[14] should be interpreted as a decision about whether, given the stakes at issue, uncertainty in the statistical methods employed justified the outcome. The problems they discuss relating to uncertainty arising from reference classes, and with regard to choice of statistical methods, are precisely the same limitations we see with the models employed by Nance and Morris and by Finkelstein and Levin. However, we do not accept Colyvan and Regan's characterisation of us as 'pessimistic' about the use of formal methods, nor do we think that their employment of a distinction between ambiguity and vagueness captures the thrust of our concerns about formal analysis.

Our statement that 'inferential problems at trial appear to defy formal treatment' was not meant in a pessimistic spirit. As noted above, the use of such methods may have a place in evaluating inferences for parties, witnesses and fact-finders. Our concern is with the limits that must be placed on such uses; the results of formal analysis will always be defeasible and thus cannot be equated with the true or correct value of items of evidence. This conclusion is warranted regardless of whether the models are derived from classical probability theory or the various unconventional probability theories to which Colyvan and Regan refer. Defeasibility is inescapable not only owing to potential ambiguity or vagueness with evidence (or both simultaneously), but also because every trial involves a unique combination of evidence and underlying historical events.

14 *United States* v *Shonubi*, 962 F Supp 370 (EDNY 1997); *United States* v *Shonubi*, 103 F 3d 1085 (2d Cir. 1997); *United States* v *Shonubi*, 895 F Supp 460 (EDNY 1995); *United States* v *Shonubi*, 998 F 2d 84 (2d Cir. 1993); *United States* v *Shonubi*, 802 F Supp 859 (EDNY 1992).

Finally, Colyvan and Regan's discussion of vagueness in legal concepts points toward another issue in need of clarification.[15] Much of the factual subject-matter of law involves determining what happened; other questions involve evaluative, often vague, terms such as reasonableness, negligence and obscenity. Perhaps vague probability notions could shed some light on these issues, but, to be clear, neither type of issue is reducible to formal analysis in the sense of definitively specifying the 'true' strength of the evidence.

Rhee

Rhee explores the problem of reference classes in the context of evidentiary weight and inductive (or Baconian) probability. He usefully draws attention to the fact that the limitations we discussed in the evaluation of trial evidence also arise in the context of the economic literature on settlement and bargaining. We agree with much of Rhee's paper, but we take issue with two points. First, the reference class problem remains even when evaluating the weight of the evidence or evaluating weightier evidence. Secondly, we are less reluctant than Rhee to allow parties to introduce, and fact-finders to rely on, probabilistic evidence (perhaps Rhee is a more appropriate target for Nance and for Colyvan and Regan).

Rhee focuses on the familiar gatecrasher paradox to illustrate the difference between probability and weight.[16] He relies on Keynes's description that probability measures the difference between a proposition and its negation while weight measures the amount of evidence supporting each proposition.[17] Although the mathematical probability in the example favours the plaintiff, the evidence supporting this conclusion is not particularly weighty. If we knew more about the rodeo attendees (for example, whether they were boy-scouts, their age, and so on), the ratios would change, perhaps dramatically in favour of—or against—certain defendants. Learning more information about the attendees may make for weightier evidence, but there is still no guarantee that these weightier probability assessments are taking one closer to or further away from the truth. Even if a particular defendant falls in the class of boy-scouts, for example, he may also fall into the sub-category of highly dishonest boy-scouts. This is another way of saying that narrowing the size of the class or making more nuanced probability assessments based on more information will not necessarily make for more accurate judgments. Suppose we know that the number of gatecrashers in Section A is only four out of 50, which appears to favour each defendant in this section. Still, this additional information will not necessarily constitute better evidence than our

15 See e.g. their discussion of 'more reliable'.

16 L. Jonathan Cohen, *The Probable and the Provable* (Clarendon: Oxford, 1977) 74–81.

17 John Maynard Keynes, *A Treatise on Probability* (Macmillan: London, 1920) 77.

original probabilistic evidence with regard to potential defendants in this section. Perhaps four out of the five in the first row of Section A are indeed gatecrashers (now favouring the plaintiff with regard to these defendants). What matters is whether the chosen class is homogenous. Absent this knowledge, the problem remains even for weightier evidence. So we agree with Rhee that evidentiary weight may be affected by reference classes, but weightier probability judgments are still not necessarily less prone to the problem.

Similar to Colyvan and Regan's discussion of *Shonubi*, Rhee is concerned about the potential threat to individualised justice posed by aggregate probability judgments, and he argues that reluctance to introduce mathematical methods into law appears justified on this basis. As noted above, however, we are not necessarily against the use of probabilistic methods in court. We think formal methods may be useful so long as their limits are recognised. For example, in a racial-discrimination lawsuit alleging that a plaintiff was not hired on grounds of race, data about other successful and unsuccessful applicants would be crucial evidence, just as would epidemiological studies in a tort case showing a significant increase in disease among those exposed to a particular product. There would be reference class issues here, to be sure, and this is why formal analysis by itself cannot be taken as the strength of the evidence. But explanatory relationships with the claims of the parties may qualify statistical information as important items of evidence: a discriminatory motive may best explain the hiring data; that the product causes disease may best explain an increase in disease among its users. Frequency data and formal analysis are useful to the extent that they serve these explanatory considerations, and so long as this basic constraint is recognised and respected we have no objection to their use. We turn now to consider more directly, in our reply to Laudan, the explanatory considerations that we see as fundamental to the juridical proof process.

Laudan

In our original article we contended that explanatory considerations better explain the legal proof process than probabilistic alternatives. In particular, we advanced this proposition in relation to micro-level issues such as relevance and probative value and with regard to macro-level issues involving the legal standards of proof. Laudan appears to accept our criticisms of the probability approaches,[18] but he disagrees with our appeal to explanatory considerations

18 See Larry Laudan, *Truth, Error, and Criminal Law: An Essay in Legal Epistemology* (Cambridge University Press: Cambridge, 2006) xii.

pertaining to the macro-level issues, in particular the criminal standard of proof beyond a reasonable doubt.[19]

We had used the phrase 'inference to the best explanation' (IBE) as shorthand for the idea that explanatory considerations guide inferences during the proof process and that these considerations therefore themselves explain the legal phenomena under discussion. Such considerations employ at root the idea that the better a proposition explains the relevant evidence, the more likely it is to be true. And that of two or more competing explanations, the one that better explains the evidence is more likely than the other(s) to be true. It seems to us that Laudan is attacking a different target than what we had in mind. In Laudan's words, IBE 'aims to be a universal standard of credibility'[20] and 'about what it is generally rational to believe'.[21] There may be significant philosophical interest in establishing such an account, but we had nothing so ambitious in mind. Our concerns are local and specific, restricted to the operation of the legal system and how explanatory considerations illuminate the forensic process.[22] We will now expand on what we had in mind,[23] and respond to some of Laudan's specific points along the way.

The explanation-based inferential process occurs in two stages: first generating potential explanations, and then selecting one (as the 'best' or 'better') based on explanatory criteria. The stages vary according to context. Applicable substantive law and explanations offered by the opposing party will affect parties' choices. Those with the burden of proof will select potential explanations that incorporate the formal elements they must prove; opposing parties will select explanations that exclude one or more of these elements. The criteria for selecting which explanations are better or worse will likewise vary with context. Theorists have articulated several general considerations—simplicity, consistency, coherence, consilience, and so on—some of which may be more or less relevant in any given context. As an empirical matter, jurors appear to focus on coherence, completeness and uniqueness in selecting among the various

19 He offers no view regarding the micro-level issues. Colyvan and Regan also offer some general comments regarding our appeal to explanatory considerations; our reply in this section also responds to their comments regarding IBE.

20 Laudan, 'Inference to the Best Explanation and the Criminal Standard of Proof', above at 292.

21 Ibid. at 306.

22 Also, for this reason, certain objections to IBE raised in the philosophy of science are inapplicable when applied to the legal context. For example, the law is typically not concerned with proving the existence of unobservable entities: see Larry Laudan, 'A Confutation of Convergent Realism' (1981) 48 *Philosophy of Science* 19, 33; Bas C. Van Fraassen, *The Scientific Image* (OUP: New York, 1980) 134–57.

23 For a more detailed account see Pardo and Allen, above n. 2.

explanations put forward by competing parties.[24] Other criteria may be more important in other contexts, such as consilience and convergence in science. Also important in the juridical proof context is that explanations are contrastive—they will diverge at key junctures and will be better or worse at explaining key pieces of evidence.

Turning to legal standards of proof, in civil cases fact-finders ought to infer the best explanation (and find for the party whom it favours) from the competing explanations offered by the parties or additional explanations fact-finders construct for themselves (individually or collectively during deliberation). Laudan, however, claims that IBE cannot explain the civil preponderance standard because IBE requires not just the better of the competing explanations but some additional threshold level of plausibility to count as best. We do not recognise any such legal requirement and do not endorse this model as explaining the civil standard. If both parties offer poor explanations, jurors should find for the better of the two parties' explanations (unless they can construct a better one for themselves, based on the evidence they have heard). If the explanations are so bad (or good) that the jurors cannot decide between them, then the decision should go against the party with the burden of persuasion on the issue. If this is not required by a universal model of IBE, so be it. We are concerned with the role of explanations guiding inferences in legal fact-finding; the better the explanation, the more likely it is to be true.

Things get more complicated in the criminal context, which is Laudan's main focus. Laudan identifies two main problems for IBE in explaining the criminal standard. First, the 'best of a bad lot' problem recognises that if jurors infer the best explanation from a pair (or more) of lousy explanations put forward by the prosecution and the defence, this would, contrary to normal expectations, convict too many defendants whom, we recognise intuitively, have not been proven guilty beyond a reasonable doubt. Secondly, the 'best of a good lot' problem posits that if jurors infer the best explanation from comparatively strong competing narratives put forward by each side, this would again convict defendants on less than proof beyond a reasonable doubt. We agree with these criticisms and do not endorse such a model in this context. Jurors ought to convict when there is a plausible explanation consistent with guilt and no plausible explanation consistent with innocence, and jurors ought to acquit either when there is no plausible explanation consistent with guilt or a plausible explanation consistent

24 See Nancy Pennington and Reid Hastie, 'A Cognitive Model of Juror Decision Making: The Story Model' (1991) 13 *Cardozo Law Review* 519.

with innocence.[25] Laudan offers several criticisms of this account.[26] First, he claims it is not inference to the *best* explanation anymore; in fact it involves sometimes rejecting the best explanation and acquitting. Secondly, this is no longer a general standard of credibility and is too dependent on context. Finally, the account severs IBE's link between being the best explanation and being probably true. Our reply to the first two points is the same: we agree. In this context, one is no longer inferring the best explanation of those available, and this is because the explanatory considerations shift for the purposes of criminal adjudication. To the extent that this rationalisation deviates from a universal model of inference, again so be it. On the third criticism, we agree that our account severs the link between the best explanation and being 'probably true', but this is a necessary feature of the beyond-a-reasonable-doubt standard itself. The standard, as Laudan recognises, skews errors in favour of the defendant; therefore, the explanatory requirements ought to shift as well. So long as the general relationship holds, that the better the explanation the more likely true,[27] then, for example, having a plausible explanation of innocence makes the possibility of factual innocence likely enough that the jury ought to acquit, even if this is not the *best* of the explanatory lot.

We close by returning to the micro-level issues, which were not Laudan's focus but which were discussed by other contributors to this symposium. Similar explanatory considerations inform legal conceptions of relevance and probative value. An item of evidence is relevant if it is explained by the particular explanation offered by the party tendering the evidence, assuming the explanation matters to a fact of consequence to the substantive law. The probative value of this evidence will depend on the strength of the pertinent explanation: the more it is explained, the more probative; the less it is explained, the less so. The strength of the desired inference will depend on all the other relevant evidence and any competing (contrasting) explanations. We think this schema is fundamental to understanding relevance, probative value and factual inferences in legal adjudication. All of the variations on formal analysis discussed by other contributors, like probability approaches in general, are useful precisely to the extent that they contribute to this understanding, and less so to the extent they do not.

25 Ronald J. Allen, 'Rationality, Algorithms, and Juridical Proof: A Preliminary Inquiry' (1997) 1 E & P 254.

26 Laudan, 'Inference to the Best Explanation and the Criminal Standard of Proof', above at 292.

27 See David Lewis, 'Elusive Knowledge' *Papers in Metaphysics and Epistemology* (Cambridge University Press: Cambridge, 1999) 433.