Science & Justice 59 (2019) 467–469

Contents lists available at ScienceDirect

Science & Justice

journal homepage: www.elsevier.com/locate/scijus

Letter to the Editor

Letter to the editor: Reply to Biedermann and Gittelson

Dear Editor,

Knowing that criticism is a form of flattery, I am grateful to Biedermann and Gittelson for their serious-minded critique [1] of my recent article, "Strategic choice in linear sequential unmasking" [2]. They organized their remarks under four headings, which I will address in order. I will focus on the main points and ignore relatively minor misunderstandings. Naturally, I hope Biedermann and Gittelson will forgive any misunderstandings of my own that may be reflected in this letter.

The object of study. Biedermann and Gittelson seem to doubt that descriptive decision theory is of any use in evaluating alternative proposals for improving crime labs. Whether a proposal will in fact improve things, they tell us, is an "empirical question" best addressed with "actual experiments in controlled conditions." A rigorous experiment can indeed be informative as my own work may, perhaps, suggest [3,4]. But the space of possible protocols is vast, perhaps infinite. Controlled experiments are costly and time consuming. If we had to rely on experiments alone, we could not hope for progress in our knowledge of how to improve forensic science. If people tend to notice and respond to incentives, then descriptive decision theory is a useful tool of analysis that we should not spurn. And, indeed, my central point was precisely that it "has been underutilized." See Whitman and Koppl [5] for another example of descriptive decision theory applied to forensic science.

Biedermann and Gittelson think it was unclear whether my paper was something about "allele designation" or something about "protocol prescriptions." An author must always take the blame for any confusion or misunderstanding about their meaning, and so I must apologize for being unclear. My main point was, again, that descriptive decision theory "has been underutilized." My critique of linear sequential unmasking was meant to illustrate the value of decision theory by drawing substantive conclusions about one proposed "protocol prescription", namely, confidence attributions in linear sequential unmaking.

Modeling allele designation decisions. Biedermann and Gittelson's central point under this heading seems to be that a forensic scientist "should be concerned about the accuracy of his or her statement," rather than how it will be received by others. I agree. They *should* be above it all, but *will* they? Divine creatures can fully transcend their incentives, humans cannot. Humans tend to respond to subjectively perceived costs and benefits, even when they should not respond to such incentives. They respond to incentives even when they don't want to. This responsiveness may exist at an unconscious level, as when motivated examiners see what they want to see rather than what is there. (Recall that the "three keys to observer effects in forensic science" are "the analyst's state of *expectation*, the analyst's state of *desire*, and the degree of *ambiguity* in the evidence being examined" [6].) If we want to find protocol prescriptions that will improve crime labs, we had better recognize that forensic scientists are human beings and not divine creatures. We need protocol prescriptions suitable for ordinary, fallible, striving human beings in their stumbling and bumbling. Descriptive decision theory can help us to find such human-compatible protocol prescriptions.

Biedermann and Gittelson seem to agree that humans respond to incentives when they call for a "scoring rule" that "encourages the elicitation of sincere probability assertions." They say, "We argue that this is the incentive that shall drive the scientist's assessment." I confess that I am unsure what they mean by "shall." We should *instruct* the forensic scientist to respond to the scoring rule and only the scoring rule? Or we should *induce* the forensic scientist to respond to the scoring rule and only the scoring rule? If forensic scientists are like other humans, they respond to incentives. In that case, it is no use to *inform* them of what incentives we *want* them to respond to. We must *induce* appropriate behavior by structuring their *actual incentives* appropriately, which is why I have called for "well-designed error tolerance regimes" in crime labs.

Biedermann and Gittelson seem unimpressed with my call for well-designed error tolerance regimes. And yet they propose a scoring rule, Brier's "verification system" for weather forecasters, that was originally designed for that very purpose [7]. Brier notes that older scoring rules created a "difficulty" for the forecaster, who "may often find himself in the position of choosing to ignore the verification system or to let it do the forecasting for him by 'hedging' or 'playing the system'" [7]. My criticism of confidence attributions in linear sequential unmasking was based on the precisely parallel concern that forensic scientists may find themselves having to choose between ignoring their incentives and playing the system. However optimistic we may be about the ability of some forensic scientists to ignore their incentives, others will be driven to hedge and play the system. For this reason, the confidence attributions tentatively proposed by Dror et al. [8] may go wrong. We need to be sure crime lab systems induce appropriate incentives. Descriptive decision theory can help.

Brier's rule was devised for a context in which ground truth is revealed soon and with effective certainty. By Thursday we will know if Tuesday's prediction of rain Wednesday was correct. In forensic science, ground truth typically remains buried. We rarely know for sure whether the latent and rolled prints came from the same finger. Thus, it is not a simple or straightforward matter to devise a good error tolerance regime that uses Brier's rule. Who would make the ex post determinations of truth? What are *their* incentives? When would these determinations be made? By what criteria? And so on. In contrast with real case work, we do know ground truth in proficiency tests. It thus seems worth investigating whether to combine blind proficiency tests with Brier's rule to get both improved incentives for forensic scientists and reasonable estimates of a lab's epistemic performance. Whatever the merits of that tentative proposal, the general notion of quadratic scoring is suggestive, and I thank Biedermann and Gittelson for bringing it forward in their letter.

The notion of error. Biederman and Gittelson criticize me for saying,

https://doi.org/10.1016/j.scijus.2019.02.008
Received 9 February 2019; Accepted 24 February 2019

"Information revealed subsequently may induce the examiner to change their determinations. Such a change could be construed as showing that the earlier determination was an error." They chide me for neglecting the role of priors in Bayesian updating and, more importantly, for failing to recognize that "a given probability is a function of the information available at the time when the probability is asserted." (They view probabilities as epistemic, not ontological.) Here again, I must take my responsibility as an author and apologize for being unclear. I did not mean to say that changing a determination on the basis of new information does *in fact* show that one had previously erred. It may in some cases, but not in others. Rather, there is a risk (to the forensic examiner) that such a revision may be so *construed*, even when it is inappropriate to do so. Indeed, I was trying to say that we should not want forensic scientists to be blamed and shamed for supposed "errors." We should want an "error tolerant" regime. "But," I said, "an organization or system cannot be error tolerant if those who make errors are afraid of being caught in an error from fear of blame, scorn, dismissal or the like."

Descriptive capacity of normative analyses. Under this fourth and final heading, Biedermann and Gittelson insist that normative analysis can have descriptive content. "[I]t cannot be denied that normative decision theory is descriptive with respect to the *problem* faced by scientists in the first place, that is the decision to be made under uncertainty – and its fundamental ingredients." I deny it. The decision problem faced by a forensic scientist is framed by the scientist and not by Professors Biedermann and Gittelson. Perhaps we should regret this fact. Forensic science might be better if forensic scientists would toss out their actual fears and desires and respond only to the sort of incentives Biedermann and Gittelson favor. But that's not what happens. People respond to their subjectively perceived costs and benefits, whatever they may be. Each individual scientist frames their decision problem for themself, whether we like it or not. Were it not so, we would not need Brier scoring. We would not need anything that "encourages the elicitation of sincere probability assertions" from forensic scientists. That Biedermann and Gittelson call for such a system suggests that they are not being consistent about whether actual decisions are framed by the actual decision makers or by third parties.

My point about who frames the decision has at least two aspects. First, the forensic scientist's theory may not be current and correct. My article has the example of crazed glass as an spurious indicator of rapidly increasing temperature. Second, the forensic scientist's preferences ("utility") may not be consistent with scientific objectivity. My article has the example of wishing to vindicate the police theory in a case. Even apart from such problems, normative decision theory as proposed by Biedermann and Gittelslon would seem to make forensic science unreliable in a precise technical sense.

The word "reliability" has many meanings. One standard meaning that occurs frequently in forensic science holds that a forensic technique is "reliable" if different persons get similar results when applying the technique. "In statistics and science in general validity is synonymous with accuracy, and reliability is synonymous with precision" [9]. As Kaye and Freedman explain, "reliability refers to reproducibility of results" [10]. A reliable technique will have a low "within-observer variability" and a low "between-observer variability" [10]. It seems reasonable to ask that forensic scientists use reliable, reproducible techniques. As I noted in my article, Biedermann, Gittelson and their co-authors say, "the scientist's preferences will vary from one scientist to another, and may also vary from one case to another in function of the severity of the case" [11]. Such variation renders their proposed techniques correspondingly unreliable and irreproducible. To the extent that the results of forensic analysis vary from one case to another because "the scientist's preferences will vary from one scientist to another," the law is being applied differently to different persons. To that extent, the principle of equal protection of the laws is compromised. To the extent that the results of forensic analysis vary from one case to another "in function of the severity of the case," the effective standard of proof will depend in part on the preferences of forensic scientists rather than the legal principles of "proof beyond a reasonable doubt" and, for civil cases, "preponderance of the evidence."

Biedermann, Gittelson, and their co-authors do not consistently represent the forensic scientist's preferences as personal and subjective. They sometimes say, the "utility function is imposed by the objectives and preferences of society" [12]. I wonder which is their current position: preferences come from the scientist or preferences come from society? In any event, they might invoke this idea of the "preferences of society" to get around the problem of low reliability. But doing so would raise its own problems. What mechanism could ensure that the forensic scientist would act on the "preferences of society" rather than their own subjective preferences? Here again, *instructing* them to do so is ineffective. They must be *induced* to do so. Until it can be shown how this could be done, appeal to the "preferences of society" has no practical meaning. It is not clear that society has any preferences at all. The Nobel laureate economist Kenneth Arrow famously showed that the collective may not have a coherent set of preferences that meaningfully reflect the preferences of its members. He expressed the gist of it this way: "If we exclude the possibility of interpersonal comparisons of utility, then the only methods of passing from individual tastes to social preferences which will be satisfactory and which will be defined for a wide range of sets of individual orderings are either imposed or dictatorial" [13]. If we all have identical preferences, there is no problem. But if different people have different preferences, there may or may not be a coherent set of social preferences beyond what might be imposed or dictatorial. Even if we could somehow get around the Arrow problem, how would we determine what, precisely, the "preferences of society" are? Who would make that determination? What are their incentives? And how could we induce them to report the "preferences of society" accurately? It seems to me that the notion of "preferences of society" has no legitimate application to forensic science. In that case, however, we are brought back to the problem of low reliability that would be created by using normative decision theory in the way Biedermann, Gittelson, and others have proposed.

In their concluding remarks, Biedermann and Gittelson say, "our main difference in view regards the definition of the states of nature." They dislike the idea of forensic scientists paying attention to their actual fears and desires. They like the idea of a forensic scientists "conditioning on the true state of nature." I suppose they are right; we have different "states of nature" in our analyses. I proceed on the assumption that forensic scientists will define the "states of nature" for themselves, perhaps in ways we dislike. Biedermann and Gittelson seem content to tell forensic scientists, in effect, "Don't do that!" I prefer the "system approach" of James Reason. "The basic premise in the system approach is that humans are fallible and errors are to be expected, even in the best organisations. Errors are seen as consequences rather than causes, having their origins not so much in the perversity of human nature as in 'upstream' systemic factors" [14]. Let us take Reason's "system approach" to improving forensic science. And let us not neglect the help that descriptive decision theory can offer in the search for human-compatible protocol prescriptions.

Sincerely,

Roger Koppl.

References

[1] A. Biedermann, S. Gittelson, Letter to the editor: commentary on "Strategic choice in linear sequential unmasking, Roger Koppl", Sci. Just. (2019), https://doi.org/10.1016/j.scijus.2018.10.010.
[2] R. Koppl, Strategic choice in linear sequential unmasking, Sci. Justice 59 (2019) 166–171 (forthcoming).
[3] E.J. Cowan, R. Koppl, An experimental study of blind proficiency tests in forensic science, Rev. Austrian Econ. 24 (2011) 251–271.
[4] R. Koppl, R. Kurzban, L. Kobilinsky, Epistemics for forensics, J. Soc. Epist. 5 (2008) 141–159.
[5] G. Whitman, R. Koppl, Rational bias in forensic science, Law Prob. Risk 9 (2010) 69–90.

[6] R. Koppl, D. Charlton, I. Kornfield, D. Krane, M. Risinger, C. Robertson, et al., Do observer effects matter? A comment on Langenburg, Bochet, and Ford, Forensic Sci. Policy Manag. 6 (2015) 1–6.
[7] G.W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Rev. 78 (1950) 1–3.
[8] I. Dror, W. Thompson, C. Meissner, I. Kornfield, D. Krane, M. Saks, et al., Context management toolbox: a linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision-making, J. Forensic Sci. 60 (2015) 1111–1112.
[9] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (2011) 91–98.
[10] D.H. Kaye, D.A. Freedman, Reference guide on statistics, Reference Manual on Scientific Evidence, third edition, The National Academies Press, Washington, DC, 2011, pp. 211–302.
[11] S. Gittelson, A. Biedermann, S. Bozza, F. Taroni, Decision analysis for the genotype designation in low-template-DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 118–133.
[12] S. Gittelson, S. Bozza, A. Biedermann, F. Taroni, Decision-theoretic reflections on processing a fingermark, Forensic Sci. Int. 226 (2013) e42–e47.
[13] K. Arrow, A difficulty in the concept of social welfare, J. Pol. Econ. 58 (1950) 328–346.
[14] J. Reason, Human error: models and management, Brit. Med. J. 320 (2000) 768–770.

Roger Koppl
Department of Finance, Syracuse University, 721 University Avenue, Syracuse, NY 13244, United States
E-mail address: rkoppl@syr.edu.