Using Inferential Robustness to Establish the Security of an Evidence Claim

Kent W. Staley

Saint Louis University

Abstract

Evidence claims depend on fallible assumptions. This paper discusses inferential robustness as a strategy for justifying evidence claims in spite of this fallibility. I argue that robustness can be understood as a means of establishing the partial *security* of evidence claims. An evidence claim is secure relative to an epistemic situation if it remains true in all scenarios that are epistemically possible relative to that epistemic situation.[1]

1 Three strategies

Assessments of evidence depend on fallible assumptions of various kinds. Heinrich Hertz mistakenly assumed that his cathode tubes were sufficiently evacuated to display the deflection of cathode rays by an electric field, supposing the rays to be electrically charged. Galileo incorrectly assumed that the combined orbital and rotational motions of the earth would producing a sloshing of the oceans that would be observed as tidal phenomena, and concluded that tidal phenomena constituted evidence for such terrestrial motion. In 1984, Carlo Rubbia of the UA1 Collaboration at CERN believed that, in data collected from high energy proton collisions looking for a decay signature of the top quark, background processes had been sufficiently accounted for to regard the remaining excess of events as evidence of top quark production. A significant source of background had been overlooked, however. There was no evidence for the top quark until ten years later.[2]

One strategy for coping with the fallibility of such evidence claims is to seek further support for assumptions about which one is uncertain, and to use only those assumptions whose support is thus strengthened. Call this the *strengthening* strategy. Its strongest form would be to rely only on those assumptions for which one has *conclusive evidence*. But the effectiveness of the conclusiveness standard is in tension with its practical value. If we give "conclusive evidence" a sufficiently strong reading to remove all possibilities of error in one's background assumptions, however remote or implausible, one will rarely if ever be in a position to make interesting evidence claims.

A second strategy — call it *weakening* — is to alter one's conclusion to a claim logically weaker than the original, so that potentially false assumptions are no longer needed. Instead of claiming that one has discovered evidence for, say, the positron (in the sense of the anti-electron of current theory), one might claim only evidence for a positively charged particle with mass on the order of that of the negatively charged electron, as Carl Anderson did in his 1933 "discovery of the positron." This strategy also has an extreme version, which is to forego any ampliative inference whatever and conclude only what can be derived deductively from the data in light of the assumptions of which one is already certain.

A third strategy (used possibly in conjunction with the other two) is to appeal to robustness considerations (Campbell & Fiske 1959; Levins 1966; Staley 2004; Wimsatt 1981). In appealing to robustness, one bases one's evidence claims on the convergent outcomes of multiple tests drawing upon independent assumptions. This broad characterization of robustness runs together inferential strategies that can be distinguished both in terms of their function and their rationale. The present discussion will focus on what Woodward calls "inferential robustness," which he characterizes as arising in situations where there is a single body of data D, to be used to reach a conclusion S, where drawing such a conclusion requires the use of some additional assumption drawn from a collection of different competing possibilities A_i. If D supports S regardless of which of the A_i is used, then we will say that the support D provides for S is *inferentially robust* with respect to A_i (Woodward 2006, 219).

The aim of this paper is to articulate and defend a dimension of epistemic assessment that is appropriate for understanding these strategies, particularly inferential robustness. (Henceforth, all references to robustness should be understood in terms of inferential robustness.) Such an effort is needed because, although we might assume that some kind of requirement of *reliability*, in a broadly frequentist sense, of one's inferential procedure is relevant, if not central, to the notion of empirical evidence, there are reasons to suspect that reliability considerations are insufficient by themselves to explicate what is at stake in coping with the fallibility of background assumptions, at least insofar as these are addressed by inferential robustness. A somewhat different perspective can put the strengthening, weakening, and robustness strategies into focus, so that it can be seen clearly how all three strategies contribute in different ways to the same epistemic goal, which I articulate in terms of *security.*

In what follows, I will illustrate robustness with a recent example from the search for dark matter (section two). That example will also serve as the basis for an argument to the effect that what is at issue in at least some uses of the robustness strategy cannot be captured by appealing to reliability considerations alone (section three). In section four, I introduce as a heuristic the notion of a space of epistemic possibilities and define security. I use that definition to frame general arguments for the epistemic relevance of security in section five, and I conclude with a clarification of the robustness strategy and some comments on the relationship between security and probability.

2 An Example: Evidence for Dark Matter?

Here I discuss a recent example of the inferential robustness strategy, used in the assessment of evidence for weakly interacting massive particles (WIMP's). The evidence claim in question is that data revealing an *annual modulation* of the detection rate in particle interactions with nuclei in scintillating sodium-iodide crystals is evidence that those crystals are immersed in a 'WIMP wind' due to Earth's movement through a WIMP halo that pervades the galaxy. That claim has been put forth by the DAMA-NaI group, operating deep underground at the Gran Sasso laboratory in Italy. DAMA's claim has been contested by the negative results of other groups using more conventional experimental methods (i.e., looking for *statistical excesses* in the detection rate beyond expectations from background). The dispute has been discussed in admirable detail by Robert Hudson (2007a, 2009).[3] Here I take no stance regarding the dispute over DAMA's results, but only seek to understand the nature of one argument implicated in that dispute. Thus the present paper will only sketch very roughly the broad outlines of DAMA's analysis and result.

As Hudson explains, the initial positive result from DAMA was presented both in terms of an estimate of the WIMP mass (m_W) and interaction cross section ($\xi\sigma_p$), and in the form of a "contour" in the space of possible values of m_W and $\xi\sigma_p$. DAMA's search strategy is to examine the distribution of interaction events with regard to energy, location within the detector, and time over a one-year data collection period. This information is summarized in the number N_{ijk}, where index i indicates the i-th day, j indicates the j-th detector, and k indicates the k-th energy interval or "bin", each event being recorded as falling into a bin of width 1 keV, from 2-20 keV.

The analysis of this data assumes a *theoretical model* in which the earth moves through a "halo" of WIMP dark matter with a variable velocity

$v_r(t) = V_{Sun} + V_{Earth}\cos\gamma\cos\omega(t - t_0)$ in the galactic frame, where V_{Sun} is the Sun's velocity with respect to the halo, V_{Earth} is the Earth's orbital velocity about the sun, $\gamma = 60°$ is the angle of inclination of Earth's orbital plane with respect to the galactic plane, $\omega = 2\pi/T$ with $T = 1$ year, and $t_0 \simeq$ June 2^{nd}. This model is used to derive a first order Taylor approximation for the signal rate in the k-th energy interval as $S_k = S_{0,k} + S_{m,k}\cos\omega(t - t_0)$. Consider this, with the parameters, respectively, for the unmodulated and modulated terms, $S_{0,k}, S_{m,k} \neq 0$ for at least some k as the theoretical model for a WIMP annual modulation.

To make the connection between this theoretical model and the results mentioned above, DAMA assumes a *model of the data* as being generated by a Poisson process with mean value $\mu_{ijk} = (b_{jk} + S_{0,k} + S_{m,k}\cos\omega(t_i - t_0))M_j\mathit{\Delta} t_i\mathit{\Delta} E\epsilon_{jk}$. Here b_{jk} represents a time independent background, M_j is the mass of the j-th detector, $\mathit{\Delta} t_i$ is the actual running time for the detector on the i-th day, $\mathit{\Delta} E = 1$ keV represents the width of the energy intervals, and ϵ_{jk} is the analysis cut efficiency. A time correlation analysis is then employed that uses a maximum likelihood method to produce an estimate of the WIMP mass and interaction cross section. The likelihood function $\mathscr{L}$ can be written in terms of μ_{ijk} and N_{ijk}, and the maximum likelihood can be determined by minimizing the function $y = -2\ln(\mathscr{L}) - \text{const}$. Such an analysis, carried out on their first year of data, yields the estimates $M_W = (59^{+36}_{-19})$ GeV and $\xi\sigma_p = (1.0^{+0.1}_{-0.4})\,10^{-5}$ pb. In addition, using the ratio λ of likelihoods between the hypothesis H_1 of annual modulation, with specific values of m_W and $\xi\sigma_p$, and the hypothesis H_0 of no annual modulation, they define -2lnλ as a χ^2 statistic, and generate a plot of the region in $m_W - \xi\sigma_p$ space where H_1 is favored over H_0 at a 90% confidence level.

The 2002 paper by Belli et al. that features a robustness analysis advertises itself as an extension of the "previous analyses" (those using the just described methods of data analysis (Bernabei, Belli, Montecchia, et al. 1998, 1999; Bernabei, Belli, Cerulli, et al. 2000)) "by discussing in detail the implications of the results of the uncertainties on the dark matter galactic velocity distribution" (Belli, Cerulli, Fornengo, & Scopel 2002). More specifically, those earlier analyses adopted the "standard" isothermal sphere model of the WIMP galactic halo. Belli et al. note that, in spite of its simplicity, a number of the assumptions of that model "are not strongly constrained by astrophysical observations" (ibid., 2). Moreover, the expected rate of WIMP interactions is determined in part by the distribution function for WIMPs in their six-dimensional position-velocity phase space, a function that in turn depends on the model of the galactic halo. That expected rate has consequences for the likelihood functions on which the earlier analyses depended. Thus, they seek to "study in a systematic way possible departures from the isothermal sphere model . . . specifically . . . modifications arising from the various matter density profiles, effects due to anisotropies of the velocity dispersion tensor and rotation of the galactic halo" (ibid., 1). The paper proceeds to examine four general classes of galactic halo models: spherically symmetric matter density with isotropic velocity dispersion, spherically symmetric matter density with nonisotropic velocity dispersion, axisymmetric models, and triaxial models.

Contours are presented for representatives from each class of models. The details need not further detain us. Noteworthy, however, is the conclusion drawn: "The hypothesis of WIMP annual modulation, already favored in the previous studies by using an isothermal sphere, is *confirmed in all the investigated scenarios*, and the effects of the different halo models on the determination of the allowed maximum likelihood region in the WIMP mass and WIMP-nucleon cross section have been derived" (ibid., 16, emphasis added).

3 Reliability and Robustness

Here I argue that frequentist reliability considerations, narrowly construed, are insufficient to account for the epistemic value of the robustness strategy as employed in this case. The DAMA group offers their robustness analysis of their evidence claim in order to explain "the implications on [their previous results] of the uncertainties on the dark matter galactic velocity distribution" (ibid., 1). That is, they had previously employed a model of the galactic dark matter halo that, supposing there is such a halo, might not correctly describe it.

Might the correct way to understand their analysis be that they are attempting to defend the reliability of their results, by employing a procedure that restricts the probability of arriving at an erroneous result? For example, we might view them as applying a *severe-test requirement*, in Mayo's sense (Mayo 1996), to their earlier evidence claim. That requirement can be framed as follows: Suppose that hypothesis H is subjected to test procedure T, resulting in data x_0; then H's passing T with x_0 constitutes the passing of a severe test (and hence evidence for H) just in case x_0 fit H, and the probability of H passing T with an outcome such as x_0 (i.e., one that fits H at least as well as x_0 does), given that H is false, is very low (ibid., esp. 178–87).

DAMA's original evidence claim does indeed seem to rest on the satisfaction of the severe test requirement. Their use of the confidence level construction method makes this a natural construal, as the rationale for that method depends in part on the long-run error characteristics ensured by the appropriate use of a test statistic that follows a χ^2 distribution. However, the robustness argument offered by Belli et al. eludes such a characterization in terms of severity or even some more general frequentist notion of reliability. The paper seeks to address an uncertainty regarding an assumption about the galactic halo that is used in defending the reliability of their original inference (from the annual modulation data to the contour described above). However, the argument is obviously not meant to give evidence that the original assumption regarding the galactic halo is true, since the paper discusses other possible models and makes no effort to argue against them. To put the point again in terms of severity, the robustness argument does not itself apply the severe test requirement to any hypothesis, but rather seems directed at possibilities of error that would undermine DAMA's claim to have severely tested (and passed) the hypothesis of WIMP annual modulation.

That suggests the possibility that DAMA is here attempting to give *second-order evidence* (Staley 2004): they are giving evidence, based on the agreement between the contours generated by different galactic halo model assumptions, that the annual modulation data really are evidence for the existence of WIMP dark-matter. On the severe-testing account, this would require showing that, assuming the annual modulation data is not evidence for a WIMP annual modulation, there is a very low probability these different analyses would yield contours that agree as well as these do. However, nowhere in the paper presenting this analysis can such an argument be found. Indeed, it is difficult to see how such an argument *could* be made within the domain of frequentist statistics. To do so would require answering the difficult question: On what would the error rates of such a test depend? More precisely, how would one model scenarios in which the annual modulation data are not evidence for a WIMP annual modulation so as to be able to estimate such error rates, even qualitatively?[4]

I propose that a more satisfactory understanding of Belli et al.'s reasoning can be achieved if we attend more closely to the kind of problem that this kind of robustness argument seeks to address and what is distinctive about its approach to that kind of problem. Specifically, this type of robustness argument responds to uncertainty regarding assumptions, and does so, not by removing that uncertainty (as in the strengthening strategy), but by showing how evidence claims remain valid in spite of that uncertainty. In the next section, I begin the articulation of a framework for making sense of such a strategy.

4 The space of epistemic possibilities and security

When an investigator puts forth an empirical evidence claim, she relies on a number of other claims. Some of these are claims may be known, while others might be relied upon without being known to be true.

The idea of a space of epistemic possibility gives us a way of picturing this situation in terms of a space of scenarios that might *for all we know* be actual. In his seminal formulation Hintikka (1962) takes expressions of the form "It is possible, for all that S knows, that P" to have the same meaning as "It does not follow from what S knows that not-P." This account has been contested, however (see, e.g., Kratzer 1977; DeRose 1991; Chalmers 2011), and attempts to provide semantics for epistemic modals now focus more broadly on the conditions under which one may correctly assert "It might be that P" and related statements. As far as I can tell, nothing in the discussion that follows turns on just which analysis we use. It should, however, be emphasized that on no account of epistemic modals is it the case that merely believing that P is false makes it false the P might be true.

For purposes of the present discussion, what is crucial about the space of epistemically possible scenarios is that as knowledge is gained, more scenarios are ruled out, and the space of what is epistemically possible shrinks (Chalmers 2011). To state it with a little more precision, though still informally: If (i) Ω is the space of epistemically possible scenarios relative to a body of knowledge K, (ii) Ω' is the space of epistemically possible scenarios relative to K', and (iii) $K \subset K'$, then

$\Omega' \subset \Omega$.

With this heuristic picture in mind, the investigator who seeks to make an evidence claim can be thought of as not knowing 'where she is' in the space of epistemic possibilities relative to her knowledge. Of course, the assumptions she uses in advancing her evidence claim that she *knows* to be true, will be true throughout that space, but other claims that she uses may be true in some regions and false in others. What is more, the evidence claim itself may or may not be true throughout the entire range of epistemic possibilities.

In practice, this problem is addressed by investigators as they try to anticipate objections that they might encounter in the presentation of conclusions from experimental data. Many such potential challenges can be thought of as presenting possible scenarios in which the experimenters have gone wrong in drawing the conclusions that they do (what we might call *error scenarios*). Such challenges are not posed arbitrarily or simply on the grounds that they are logically possible. Rather, both experimenters in anticipating challenges and their audience in posing them draw upon a body of knowledge in determining the kinds of challenges that are significant (Staley 2008).

Security is here proposed as a heuristic that might help systematize the strategies that experimenters use in responding to a generic problem: the investigator has good reason to consider *whether it is possible* that her evidence claim is false. If it is possible, what is the range of possible scenarios in which it is false, and can steps be taken to eliminate some of those possibilities? I propose that the following working definition reflects certain key elements of the problem that faces investigators justifying evidence claims based on fallible assumptions, such that it may prove a useful starting point for more systematic work to come.

JE (Security)**.** *Suppose that Ω_0 is the set of all epistemically possible scenarios relative to epistemic situation K, and $\Omega_1 \subseteq \Omega_0$. A proposition P is* secure throughout Ω_1 *relative to K iff for any scenario $\omega \in \Omega_1$, P is true. If P is secure throughout Ω_0 then it is* fully secure*.*

The notion of an epistemic situation is borrowed from Achinstein (2001), who describes an epistemic situation as a situation in which "among other things, one knows or believes that certain propositions are true, one is not in a position to know or believe that others are, and one knows (or does not know) how to reason from the former to the hypothesis" (ibid., 20).

Here I wish to focus on applying the concept of security to claims about evidence – i.e., claims taking the form 'Data x_0 are evidence for the hypothesis that H.' An evidence claim is secure for an agent to the extent that it holds true across a range of scenarios that are epistemically possible for that agent. Exactly which scenarios are epistemically possible for a given agent is opaque, and not all epistemically possible scenarios are equally relevant, so the methodologically significant questions turns out to be centered on *relative security*: how do investigators make their evidential inferences more secure? And which scenarios are the ones against which they ought to secure such inferences? (Consequently, there is no need to settle on any particular analysis of knowledge, since neither the enhancement of relative security nor the establishment of security across a range of possibilities specified as relevant requires one at any point to identify exactly what it is that a person knows.)

5 A general argument for security

Before showing how security helps us to understand the robustness strategy, I wish to give a general defense of the epistemic relevance of security by showing that, all else being equal, an investigator should prefer, on error-avoidance grounds, making an evidence claim that is *more* secure to making a claim that is less secure.

Suppose that S is a person engaged in empirical inquiry with regard to some question Q, to which H is a hypothetical answer. Suppose that x_0 represents a possible body of data relevant to H. C is the claim "x_0 is evidence for H." K represents S's epistemic situation at a particular time, Ω_0 is the set of all scenarios epistemically possible relative to K, and C is secure across $\Omega_1 \subset \Omega_0$. Suppose further that S is error-averse in the sense that she seeks to make evidence claims that will not be refuted by subsequent inquiry.

Now suppose that there is a strategy available to S that results in S being able to make an evidence claim C' that is secure across Ω_2, where $\Omega_1 \subset \Omega_2$. This might be accomplished by replacing H with a logically weaker H' (weakening strategy) or by gathering additional information that supports auxiliary assumptions (strengthening). Either way, there is an asymmetry between the potential failures of the two claims C and C'. For any epistemically possible scenario in which C' fails, C also fails. But there are some epistemically-possible scenarios in which C fails, but C' does not. Thus the scenarios in which C' continues to be upheld as true include all those in which C continues to be upheld as true, as well as some of those in which C is discovered to be false. Since S aims to avoid making evidence claims that are discovered to be false, she should prefer, all else being equal, to make claim C' rather than C. I hasten to note, however, that the very fact that a strategy is required to make the claim C' rather than C ensures that not all will be equal. Pursuit of either weakening or strengthening strategies will involve some cost, which is relevant to the choice of epistemic strategy. The point of this argument is not to show that one should always pursue a security-enhancing strategy, but only to point out the epistemic *relevance* of security considerations to decisions about the handling of evidence.

It will be noted that, as described, the secure regions for claims C and C' were nested, and one might wonder what might be said about cases in which this is not so. For example Ω_1 and Ω_2 might overlap without either being contained in the other, or they might be entirely disjoint. It would be tempting here to say that one should then consider the size of the secure regions of the two claims in deciding what claim would be preferable. Indeed, this seems to be the right answer, provided that the crucial condition "all else being equal" were understood in the sense of "there is no more reason to think that one is in one region of the space of epistemic possibility than in any other." However, this latter condition is rarely, perhaps never, satisfied. Although one is aware in an evidential inference that there is a range of possibilities of error, some of these possibilities will be more worrisome than others. Clearly, some means of weighing the relevance of scenarios is needed. Although I attempt no solution to this problem in the present paper, I will comment on it in the conclusion.

6 Conclusion: Robustness, dark matter, and probability

We can now revisit the robustness strategy and see more clearly how it works to establish security of evidence. First, note that robustness is distinct from weakening and strengthening strategies that *increase* the security of an evidence claim. An appeal to robustness serves to *show* that the claim C is secure across a range of possible scenarios Ω_R (represented by the different possible auxiliary assumptions), where in the absence of the appeal to robustness one might know only that C is secure across some smaller range of scenarios $\Omega \subset \Omega_R$, or be quite uncertain about the possible conditions under which C might fail to hold true. Although the possibility for error, objectively speaking, is unchanged, the investigator making claim C on the basis of a robustness analysis is in a better position to *argue* that C is not in error.

Consider how this applies to the dark matter case. The problem confronted by the argument given by Belli et al. is that, for all they know, the theoretical model of the WIMP galactic halo used by DAMA in their original analysis might be false. So what happens if some other model is the correct one? Belli et al. examine four classes of models, any of which might, for all they know, include the correct one, and they find that the WIMP annual modulation hypothesis is "confirmed in all the investigated scenarios" (Belli et al. 2002, 16). In other words, even if the isothermal sphere model is false, so long one of the investigated classes includes a model that is adequate for the purpose of estimating the annual modulation (Parker 2012), DAMA's data are still evidence for WIMPs. The effect of this argument, then, is to show, in a way that DAMA's original analysis did not, the extent to which the evidence for WIMP annual modulation is secure.

As noted above, not all possible scenarios are equally important in considering threats to the truth of a given evidence claim. Some possibilities (such as cosmic conspiracies) will be too implausible to bother with. One might object that the approach here discussed fails to deploy an obvious resource for dealing with this issue, which would be some kind of epistemic probability such as that deployed by Bayesians.

Although a full treatment of this issue goes beyond the scope of a brief essay, I will make two brief comments to justify my pursuit of a non-probabilistic approach.

First, I have sought to develop an account that is compatible with theories of evidence that treat probability in frequentist terms, such as Mayo's error-statistical approach (Mayo 1996; Mayo & Spanos 2009). An important motivation for advocates of error-statistical approaches is to avoid assigning probabilities to anything that resists being modeled as the outcome of a stochastic process of some sort. Introducing a probability measure over scenarios, which cannot be thus modeled (worlds not being "as plentiful as blackberries," to use Peirce's memorable phrase) would be conceptually at odds with the approach to evidence that is here assumed as a starting point.[5] If the measure were not a probability, it would need to be provided with some other interpretation to be meaningful. Having no defensible interpretation at hand, I deem it advisable for now to eschew such a measure entirely..

Unlike the first point, the second point does not require that we assume an error-statistical viewpoint. Regardless of one's preferred approach to statistical inference, in practice one must make *assumptions*. These may be regarded either as background assumptions serving to underwrite an ampliative inference (like the role of model assumptions in the inference from data to conclusion in error-statistics) or as part of the premises (as in Howson and Urbach's treatment of Bayesianism as a deductive logic of probabilities (Howson & Urbach 2006)). As part of the justification of an inference, one must defend the appropriateness of these assumptions (so that they are not *mere*, i.e. unfounded, assumptions), and a thorough justification requires consideration of the various ways, given what one knows, that those assumptions might fail to be appropriate. The notion of 'appropriateness' here will depend on the framework and perhaps even the nature of the particular inference. For the purposes of securing evidence, the concern is with those flaws in assumptions which would result in the presumptive evidence for the conclusion turning out not to be evidence for the conclusion, or evidence of a weaker sort than was thought.

For “objective” or “reference” Bayesians, for example, the choice of a prior probability distribution is a matter of finding a distribution that is noninformative in some sense. Although this sounds like a simple criterion, it is in practice not at all simple to execute, and even specifying the appropriate sense in which a distribution should be noninformative is a matter of debate (see, e.g., Berger & Bernardo 1992). Clearly, there are ways in which one might go wrong in attempting to specify a noninformative prior, and the justification of an inference based on such a prior will need to attend to these.

It might be thought that this kind of problem is avoided by subjective Bayesians, for whom an “anything goes” approach to priors is allowed. Here, too, things are not quite as simple as they might seem. Even subjective Bayesian priors are constrained to satisfy coherence, making the possible failure of a specified prior to meet this standard a matter of consideration. This is not trivial, even in apparently simple cases. Consider David Lindley's (1993) variant of Fisher's famous “lady tasting tea” example.[6] In it, a lady who is presented as a wine-tasting authority is given pairs of glasses of wine and is asked to determine by taste alone which glass holds a Californian wine and which contains a French wine. The inference concerns the probability π that she will identify them correctly. Lindley gives for his prior distribution regarding the value of π a beta distribution on the interval $0.5 < \pi < 1$. The density function for this distribution resembles the St. Louis arch and has a peak at $\pi = 0.75$. As noted by Stephen Senn, however, this distribution is an odd choice in light of the following considerations (Senn 2001):

Either the lady's belief in her abilities is justified or it is not. If it is justified, then she will be able to make the requested identifications with probability nearly equal to one. If she is not justified, then she is just guessing and the probability of success is near one-half. Senn suggests that because she just might have "a fine palate but a poor knowledge" a small probability should be reseved for π being nearly zero (she can tell the two wines apart but consistently misidentifies their provenance).[7]

The problem here is not that Lindley's prior is *wrong*, but that if Lindley's background knowledge includes the substantive knowledge about expertise and the distinction between palate and knowledge that informs the above criticism, the prior distribution that he himself claims to represent his probability assignments to the relevant hypotheses is not coherent with his background knowledge. So prior probabilities cannot simply be written down willy-nilly. Rather, the subjective Bayesian also needs to consider the ways in which, given what she knows, her assignment of a prior might fail to be appropriate.

Another kind of security problem that affects Bayesians arises in cases where the space of possibilities across which a prior distribution is specified suddenly expands because new possibilities are thought of. Posterior probabilities that were arrived at from the earlier prior become senseless. One has to "cross these out" and start over with a new prior distribution. This kind of problem is well-known in discussions of Bayesian philosophy of science (Earman 1992, ch. 8). In situations, therefore, in which the prospects are strong for previously unconceived possibilities emerging, all Bayesian inferences must be considered as vulnerable to such non-Bayesian corrections.

The upshot of these considerations is that, without contesting that the assignment of a measure across a range of scenarios might be a suitable Bayesian technique, it is a technique that *presupposes* the kind of consideration of the range of possible scenarios that a security perspective demands. Furthermore insofar as the distribution itself constitutes an additional assumption, one can apply the security framework also to that assignment, asking for the possible ways in which it might fail to cohere with one's background beliefs or knowledge. To put the point in terms of a slogan: "possibility comes before probability."

More specifically, *possibilities of error* in judgments about evidence demand attention before judgments about probability can be taken seriously. Here I have attempted to clarify the role of inferential robustness considerations in attending to such possibilties.

References

Achinstein, P. (2001). *The book of evidence.* New York: Oxford University Press.

Belli, P., Cerulli, R., Fornengo, N., & Scopel, S. (2002). Effect of the galactic halo modeling on the DAMA-NaI annual modulation result: An extended analysis of the data for weakly interacting massive particles with a purely spin-independent coupling. *Physical Review D*, *66*, 043503.

Berger, J., & Bernardo, J. (1992). On the development of reference priors. In

J. M. Bernardo, J. O. Berger, P. Dawid, & A. Smith (Eds.), *Bayesian statistics 4* (pp. 35–60). New York: Oxford University Press.

Bernabei, R., Belli, P., Cerulli, R., et al. (2000). Search for WIMP annual modulation signature: Results from DAMA/NaI-3 and DAMA/NaI-4 and the global combined analysis. *Physics Letters B*, *480*, 23–31.

Bernabei, R., Belli, P., Montecchia, F., et al. (1998). Searching for WIMPs by the annual modulation signature. *Physics Letters B*, *424*, 195–201.

Bernabei, R., Belli, P., Montecchia, F., et al. (1999). On a further search for a yearly modulation of the rate in particle dark matter direct search. *Physics Letters B*, *450*, 448–55.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation in the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Chalmers, D. (2011). The nature of epistemic space. In A. Egan & B. Weatherson (Eds.), *Epistemic modality.* Oxford: Oxford University Press.

DeRose, K. (1991). Epistemic possibilities. *The Philosophical Review*, *100*, 581–605.

Earman, J. (1992). *Bayes or bust? a critical examination of bayesian confirmation theory.* Cambridge, MA: MIT Press.

Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions.* Ithaca, NY: Cornell University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The bayesian approach* (Third ed.). Chicago and La Salle, Illinois: Open Court.

Hudson, R. G. (2007a). Annual modulation experiments, galactic models and WIMPs. *Studies in History and Philosophy of Modern Physics*, *38*, 97–119.

Hudson, R. G. (2007b). *Robustness vs. model-independence.* Available from `http://philsci-archive.pitt.edu/archive/00003209/` (Paper presented at LSE/Pittsburgh Conference on Confirmation, Induction, and Science.)

Hudson, R. G. (2009). The methodological strategy of robustness in the context of experimental WIMP research. *Foundations of Physics*, *39*, 174–93.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, *1*, 337–55.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*, 421–31.

Lindley, D. (1984). A Bayesian lady tasting tea. In H. A. David & H. T. David (Eds.), *Statstics: An appraisal* (pp. 455–77). Ames, Iowa: Iowa State University Press.

Lindley, D. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge.* Chicago: University of Chicago Press.

Mayo, D. G., & Spanos, A. (Eds.). (2009). *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science.* New York: Cambridge University Press.

Parker, W. (2012). Scientific models and adequacy-for-purpose. *The Modern*

Schoolman, *87*.

Senn, S. (2001). Two cheers for p-values? *Journal of Epidemiology and Biostatistics*, *6*, 193-204.

Staley, K. W. (2004). Robust evidence and secure evidence claims. *Philosophy of Science*, *71*, 467–88.

Staley, K. W. (2008). Error-statistical elimination of alternative hypotheses. *Synthese*, *163*, 397–408.

Wimsatt, W. (1981). Robustness, reliability, and overdetermination. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–63). San Francisco: Jossey-Bass.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*, 219–40.

Notes

[1] I am grateful for helpful comments on earlier drafts of this paper from Jan Sprenger, Joe Salerno, and audiences at Eastern Tennessee State University, University of Frankfurt and University of Wuppertal. Conversations with Aaron Cobb were especially helpful in developing the ideas here presented.

[2] In describing these examples in this way, I am using a concept of evidence that is objective in the sense that whether such evidence obtains is a matter that is independent of what anyone believes about the data, the hypothesis in question, or the relations between them (see Achinstein 2001).

[3] Hudson (2007b) employs the debate over DAMA's results to argue *against* the methodological value of robustness. I do not propose here to dispute the arguments of that paper, but only note that under Hudson's interpretation of the term, the argument here considered does not exemplify an appeal to robustness. Following Woodward (2006), I use the term more broadly than Hudson.

[4] I am not claiming that the convergence of results from independent tests *never* constitutes a severe test of any hypothesis, but only that some uses

of such convergence cannot be thus interpreted.

[5] This point applies even if, as in the case of Lewis, one takes a realist attitude toward *possible worlds*. Even if possible worlds are in fact as plentiful as blackberries (in some sense), there is no plausible sense in which the fact that we find ourselves in one particular actual world out of all the epistemically possible worlds can be regarded as the outcome of a stochastic process.

[6] I would like to thank Stephen Senn for making me aware of this example, and the critique of Lindley's prior that follows.

[7] In a separate, more detailed discussion, Lindley notes that it would be realistic to assign a small probability to this last possibility, but keeps the peak at $\pi = 0.75$ (Lindley 1984).