Varieties of Error and Varieties of Evidence in Scientific Inference

Barbara Osimani, Jürgen Landes

Abstract

According to the Variety of Evidence Thesis items of evidence from independent lines of investigation are more confirmatory, *ceteris paribus*, than e.g. replications of analogous studies. This thesis is known to fail Bovens and Hartmann (2003), Claveau (2013). However, the results obtained by the former only concern instruments whose evidence is either fully random or perfectly reliable; instead in Claveau (2013), unreliability is modelled as deterministic bias. In both cases, the unreliable instrument delivers totally irrelevant information. We present a model which formalises both reliability, and unreliability, differently. Our instruments are either reliable, but affected by random error, or they are biased but not deterministically so.

Bovens and Hartmann's results are counter-intuitive in that in their model a long series of consistent reports from the same instrument does not raise suspicion of "too-good-to-be-true" evidence. This happens precisely because they neither contemplate the role of systematic bias, nor unavoidable random error of reliable instruments. In our model the Variety of Evidence Thesis fails as well, but the area of failure is considerably smaller than for Bovens and Hartmann (2003), Claveau (2013) and holds for (the majority of) realistic cases (that is, where biased instruments are very biased). The essential mechanism which triggers VET failure is the rate of false to true positives for the two kinds of instruments. Our emphasis is on modelling beliefs about sources of knowledge and their role in hypothesis confirmation in interaction with dimensions of evidence, such as variety and consistency.

1 Introduction

Suppose you are a general doctor who needs to prescribe a therapy for a patient with an uncommon disease that you can pretty well diagnose, about whose most cutting edge treatments you are however unaware. You make a quick search on "pubmed" or other specific search engines, and find out that a new treatment has recently been tested for this specific condition in two distinct studies. Since this is not your speciality, you do not know the people who do research in this area; how good they are at making research, or how biased they may be because of possible vested interests. Suppose the studies are of equal quality; that is, they have the same capacity to detect true positives and negatives (sensitivity and specificity). Would the two studies be more confirmatory for you with respect to the hypothesis of efficacy, if they came from the same research group, or if they came from two completely independent ones?

Or, consider a court jury involved in judging a culprit for having committed a crime. They receive two forensics regarding the same type of evidence (e.g. hair material or blood or traces of genetic material), delivering some confirmatory support to the hypothesis of guilt. Would these forensics be more confirmatory, if they came from distinct labs or from the same one?

Third example. A governmental body needs to make policy decisions regarding the five next years of fiscal regulations. Before consulting expert groups, they consult scientific databases in search for empirical data regarding various fiscal interventions in diverse countries, looking for evidence as close as possible to the historical and political situation they are dealing with. They find two relevant studies, which investigated the efficacy of a given fiscal programme and delivered evidence at the same significance level and power. Would they find the fiscal scheme more or less efficacious if the two studies came from the same source or from different sources of information?

The commonsense intuition regarding these examples is that two pieces of evidence from different sources should be more confirmatory than analogous items of evidence coming from the same one. Indeed the hunch here is that "independent" evidence is more confirmatory than items of evidence coming from epistemically connected sources; as when in court trials the consistent testimonies of witnesses, who might have agreed on their reports, because they are acquainted with each other, is considered less probabative, than consistent accounts delivered

from people who never met each other before.

In the extreme case, when consistent pieces of evidence come from the same source, they may strengthen the hypothesis that the source is biased, rather than contributing to the confirmation of the hypothesis at stake. For instance, research groups might have tacit vested interests (in terms of career or indirect reward), which would motivate them to distort the evidence in one direction rather than the other.

This intuition is captured by the "Variety of Evidence Thesis" (VET), according to which varied items of evidence jointly converging towards a given hypothesis confirm it more strongly than less varied evidence, ceteris paribus. Hence, items of evidence from independent lines of investigation are more confirmatory, ceteris paribus, than e.g. replications of analogous studies.

However, there are also other considerations that play a role in this kind of higher-order evidence problems. These are considerations relating to the source(s) evidence is coming from; that is, not only whether they are independent from one another, but also, the extent to which they can individually, and jointly, be relied on (or whether you suspect them to be inaccurate or biased in some direction). Hence, whether you deem the investigated hypothesis as more or less confirmed from consistent reports, does not only depend on whether the sources are (in)dependent from one another, but also on the estimated noise or bias associated with them.

Let's suppose that two consistent pieces of evidence have been obtained from an instrument which may be very imprecise and inaccurate (delivering a high rate of false positives and negatives) or precise but positively biased (it delivers a high rate of false and *true* positives). Wouldn't it be more confirmatory in this sort of case if, ceteris paribus, the two pieces of evidence (for instance, two forensics), came from one lab only? Since it would be highly improbable to receive two pieces of evidence pointing in the same direction from the same noisy instrument, if indeed the hypothesis were not true, this would boost the probability that the two pieces of evidence come from the biased but more precise source, and therefore increase the belief that they are more likely to be true, rather than false, positives. This sort of considerations draw our attention to the fact that the confirmatory role of prima facie evidence is integrated with higher order beliefs about evidence itself: its origin and the whole data generating system which delivers it. This paper formalises such higher order beliefs in the form of dependence

and probabilistic constraints on a Bayesian network model of scientific inference.

By drawing on the literature on the variety of evidence thesis developed in formal epistemology, we connect previous efforts to model the confirmatory role of coherent evidence to recent debates on the reliability of statistical results and, more generally, to the informative value of various experimental settings. In particular, we drawn on the work of Bovens and Hartmann (2003), and the insights added by Claveau (2013) to the picture, and generalise their results. We highlight possible ways in which the impact of "higher order evidence" (such as information or beliefs about the source(s) of the evidence and the data generating process) may be factored in in its assessment.

Bovens and Hartmann analyse the interaction of instrument reliability, consistency of evidence, and (in)dependence of the instruments. In their model, consistency of evidence has confirmatory support along two ways: by boosting directly the belief that the hypothesis is true, and by increasing the confidence in the reliability of the instrument itself. Their main results show that for certain model parameters – primarily, low prior in the reliability of the testing instruments – such confirmatory boost is greater when the reports come from the same instrument rather than from independent ones. Which runs against the Variety of Evidence Thesis. Furthermore, in their model such effect monotonically increases for a greater and greater number of reports. That is, with increasing number of reports, the confirmatory boost to the hypothesis rises more, if they all come from one single instruments, instead of coming from distinct ones. This runs against the “too-good-to-be-true evidence” intuition underpinning suspicion of bias for monotone signalling sources. This intuition stems from the awareness that any measurement device, whether a physical instrument such as an oscilloscope, a monometer, a biomarker test, or a statistical study (e.g., a clinical trial or a cohort study) is always affected by random error. Hence the observation of invariably positive (or negative) signals raises suspicions of signal interpolation.

We analysed the roots for Bovens and Hartmann’s counterintuitive results in this respect and present our own model. We show that Bovens and Hartmann’s results hinge on a specific notion of reliability (and lack thereof): their information sources are either perfectly reliable, that is, they deliver positive reports with probability 1, if the hypothesis is true, and with probability

0 if the hypothesis is false. Or, if unreliable, their probability of delivering positive results is the same whether the hypothesis holds or not: they are full randomisers, and their reports are fully disconnected from reality.[1] But, the central point is that the randomisation parameter is allowed to vary strictly between 0 and 1; therefore the unreliable instrument may deliver positive reports at any rate between 0 and 1. We explain that this is the reason why the area of VET failure increases with increasing reports.

Claveau (2013) considers another kind of unreliable instrument, namely instruments affected by systematic error (bias). Such bias is deterministic: positively biased sources deliver positive reports with probability 1, and negatively biased sources with probability 0. Instead, in our model biased instruments deliver positive reports at a higher rate than standard instruments in the field that are affected by random error only. Hence, we do not assume the unreliable source to be fully disconnected from reality, or the reliable one to be a perfect signalling device; but rather the latter to be affected by random error, and the former to be positively biased with respect to the former. In our setting, the agent is confronted with uncertainty regarding whether the evidence is coming from a source affected by a standard amount of inaccuracy, or from one which tends to be positively biased, but bears anyway some relationship to reality (its information is relevant anyway). The VET fails here as well, however the area of VET failure does not grow with increasing reports as it happens in Bovens and Hartmann (2003), moreover it is a negligible area in comparison to both Bovens and Hartmann (2003) and Claveau (2013). In our model the VET tends to fail for a high false-to-true positive ratio of the "reliable instrument" with respect to the more positively biased one, that is, when the positive reports coming from the reliable instrument have a higher probability of being false than true, relative to those coming from the biased instrument.

Our results point to a more general explanation for the sort of VET violations that we encoun-

[1] "It is as if they do not even look at the state of the world to determine whether the hypothesis is true, but rather flip a coin or cast a die, to determine whether they will provide a report to the effect that the hypothesis is true. The randomisation parameter a indicates the chance that they provide a report to the effect that the hypothesis is true." (Bovens and Hartmann 2003, p. 57).

tered in ours and Bovens and Hartmann's research: greater confirmatory boost from non-varied evidence with respect to independent evidence comes in all cases via increasing the probability that we are dealing with true rather than false positives.

The VET fails whenever consistent evidence from the same instrument, rather than from independent ones, is discriminatory as to the accuracy of the source. More generally, our analysis emphasises that VET failure in the various models presented by Bovens and Hartmann, Claveau, and ourselves relies on assumptions related to higher level beliefs about diverse dimensions of evidence and their interaction.

The present paper has two aims: a) contribute to the debate on the Variety of Evidence Thesis by testing whether the previously obtained results also obtain in a different model of scientific inference[2]; b) investigate the epistemic dynamics that develop in such settings. Furthermore, we aim to link our results to the current debate on the reliability of statistical methods in light of the so called "reproducibility crisis". The link is mostly theoretical, rather than methodological, but it sheds some light on reasons why one would prefer direct (non-varied evidence) rather than indirect replication (varied evidence) depending on one's assumptions on the instruments characteristics.

The paper is structured as follows: We next present the VET and analyse the Bovens and Hartmann model. We then present our own model (Section 3) and conclude.

2 Variety of Evidence Thesis

According to the Variety of Evidence Thesis items of evidence from independent lines of investigation are more confirmatory, *ceteris paribus*, than e.g. replications of analogous studies coming from the same source.

As remarked at (Meehl 1990, p. 111): "Any working scientist is more impressed with 2 replications in each of 6 highly dissimilar experimental contexts than he is with 12 replications of the same experiment". Philosophers have cashed out this idea in terms of varied evidence providing more support to a hypothesis, than replications of the same experiment do. For exam-

[2]The term 'model of scientific inference' has been established in Claveau and Grenier (2019), Landes et al. (2018), Bovens and Hartmann (2003) and shall be used here.

ple, (Hempel 1966, p. 34) states that "the confirmation of a hypothesis depends not only on the quantity of the favourable evidence available, but also on its variety: the greater the variety, the stronger the resulting support". (Horwich 1982, p. 77) notes that: "It is an undeniable element of scientific methodology that theories are better confirmed by a broad variety of different sorts of evidence than by a narrow and repetitive set of data".[3]

This commonsensical intuition has been formally given justice in Bayesian epistemology by Fitelson showing that "two pieces of independent confirmatory evidence will always provide stronger confirmation than either one of them provided individually". Earman, p. 77-79 (see also Franklin and Howson (1984)) shows that the increment in confirmation to the tested hypotheses decreases marginally as more and more pieces of confirmatory evidence accumulate, from submitting the hypothesis to the same identical test over and over again.

However, replicated evidence is very important to the scientific enterprise (Munafò et al. (2017), Marsman et al. (2017), Romero (2016), Gelman (2015), Etz and Vandekerckhove (2016), Stanley and Spence (2014)). This stems from the awareness that measurements can be affected by random and systematic error (bias), and hence deliver false positives and negatives. In the former case, these are due to contingent influences, which average out in the long run. Instead, biases distort the evidence systematically in one direction or the other. In this case, errors do not average out, but rather reinforce across replications, and deceptively inflate accuracy in repeated measurements.[4]

More precisely, exactly with reference to these two kinds of error, there is a sort of division of

[3]In the same passage quoted above, (Meehl 1990, p. 113) goes on to "postulate a stochastic connection between the degree of evidentiary support, the number, variety, and stringency of empirical tests that the theory has passed or failed, and its verisimilitude, its closeness to objective reality". Implicit in all these considerations is obviously also the idea that the varied evidence "converges" towards the same hypothesis in the sample space. Therefore, the variety of evidence thesis can be fruitfully connected to the debate on the truth-conduciveness of coherence (a paradigmatic illustration for this connection is McGrew's distinction between *theoretical* and *evidential* consilience McGrew (2003)).

[4]This consideration dovetails with recent simulation studies showing the impact of bias on cumulative data analysis in terms of increased speed of convergence and "artificial accuracy"

labour between exact replication, and robustness studies (that is studies which test the hypothesis under different conditions or with different instruments etc.). The former aim to exclude or reduce the probability that the original result has been produced by chance (random error); the other kind of study instead aims to exclude or reduce the probability of a study artefact. By decreasing the probability of either error (or both) having occurred, one automatically increases the probability of the hypothesis being true (see also Meehl (1990) – in line with Lakatos' view of scientific progress).

The general intuition is however that the marginal informative value of exact replications decreases with their number. Hence, "everything else being the same", cost-efficiency considerations would recommend to proceed to test a different line of evidence, once one has already been confirmed, instead of insisting on obtaining the very same evidence again and again. The following formal models define how "everything else being the same" should be defined in these settings, and what these "ceteris paribus" conditions turn out to be.

2.1 The Levels Approach

The notion of evidential variety is not only related to independent testable consequences of theoretical hypotheses, but also akin to that of "robustness" Wimsatt (1981; 2012), "independent determinations" (e.g., (Weber 2005, 281-287)), and "triangulation" Heesen et al. (2019). This idea of variety pertains to possible diverse routes through which evidence for the same observable consequence of a hypothesis may be obtained, and adds a further level of variety: not only variety of conceivable consequences of the hypothesis to be tested, but also variety of instruments[5] put into place to detect them.

This intuition has been modelled in Bovens and Hartmann (2002; 2003), where variety relates both to the diverse possible observable consequences of the hypothesis at hand, and at a "lower level" of the inferential pathway, to the different experiments that may be conducted to test each of such consequences. At this lower level, variety means that different experiments are conducted with different *independent* instruments (see (Bovens and Hartmann 2003, p. 105)).

(Romero (2016), see also Stanley and Spence (2014)).

[5]Where "instruments", can be conveniently considered a placeholder for research teams, study design, experimental method and methodological assumptions etc.

2.1.1 Bovens and Hartmann's Model

The Bayesian network model of Bovens and Hartmann (2003) (see Darwiche (2009), Neapolitan (2003) for introductions to Bayesian networks) represents the hypothesis, (some of) its observable consequences, reports on whether these consequences were born out in experiments, and the reliability of instruments used in these experiments. The graph structure of the Bayesian network represents conditional independencies between the variables. Conditional probabilities attach to every variable which specify the probability of a variable given its parents.

The following binary propositional variables are used: A variable HYP where the intended meaning for Hyp is that "the hypothesis is true", similarly for variables Con_i ("consequence i holds"), Rep_i ("consequence i is reported")[6] and Rel_i (the source of the report i is reliable"), cf. (Bovens and Hartmann 2003, p. 89). REL is a binary variable, hence ρ is not the degree to which the source is reliable, but rather the probability that it is reliable, and $\bar{\rho}$ the probability that it is not reliable. A prior probability function P, defined over the algebra generated by these variables, is selected. Naturally, P is constrained to respect the conditional independencies encoded by the graph $\mathcal{G}$. Updating the prior P, by conditionalising, then allows Bovens & Hartmann to calculate posterior probabilities given experimental results.

The set of meaningful conditional probabilistic independencies in the prior P can be read off by means of the graphical d-separation criterion Pearl (2000). The graph $\mathcal{G}$ in Figure 1 depicts the situation for one single consequence. More general cases are depicted in Figure 2. These conditional independencies – denoted by $\perp\!\!\!\perp$ – are

$$HYP \perp\!\!\!\perp REL_i \quad \text{for all } i \tag{2.1}$$

$$CON_i \perp\!\!\!\perp REL_i \,|\, HYP \quad \text{for all } i \tag{2.2}$$

$$REP_i \perp\!\!\!\perp HYP \,|\, REL_i, CON_i \quad \text{for all } i \tag{2.3}$$

$$\{CON_i,\, REL_i\, REP_i\} \perp\!\!\!\perp \bigcup_{k \neq i} \{CON_k,\, REL_k,\, REP_k\} \,|\, HYP \ . \tag{2.4}$$

[6] $\neg Rep_i$ means that "not consequence i" is reported, rather than "consequence i" is not reported, i.e. the situation is one of negative evidence rather than of absence of evidence

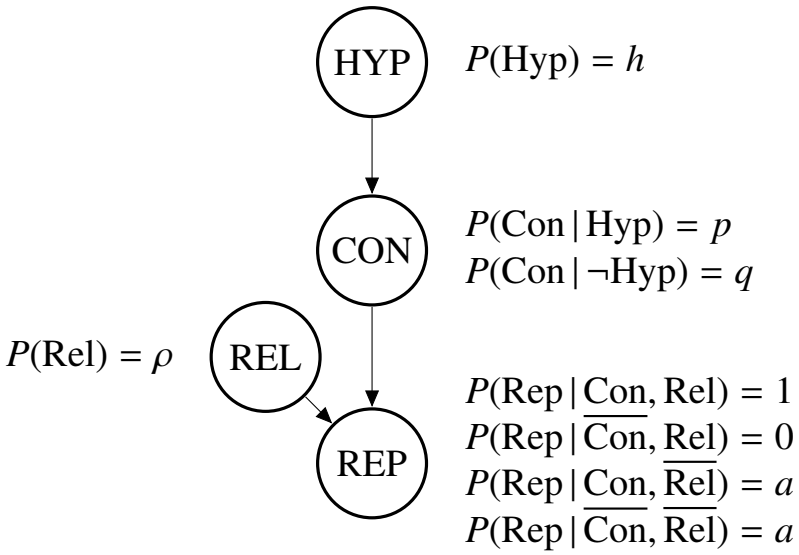

Figure 1: Hypothesis testing in the Bayesian framework of Bovens & Hartmann for one single testable consequence. All necessary parameters of the prior probability are displayed in terms of lower-case letters.

The choice of prior is further constrained by

$$P(Con_i \mid Hyp) = p_i > q_i = P(Con_i \mid \overline{Hyp}) \tag{2.5}$$

$$P(Rep_i \mid Con_i, \overline{Rel_i}) = P(Rep_i \mid \overline{Con_i}, \overline{Rel_i}) = a_i \tag{2.6}$$

$$P(Rep_i \mid Con_i, Rel_i) = 1 \tag{2.7}$$

$$P(Rep_i \mid \overline{Con_i}, Rel_i) = 0 \ . \tag{2.8}$$

(Bovens and Hartmann 2003, p. 90) take (2.5) to be their definition of what it means to be an observable consequence of a given hypothesis: it should be more probable to observe a consequence of a hypothesis, when the latter holds than when it does not.[7] (2.6) models the epistemic dynamics associated with a ransomising instrument, whose probability of delivering a report is the same (parametrised by a_i) whether the consequence holds or not. When the instrument is fully reliable, the probability of receiving a report that the consequence has been observed equals one, if the consequence holds (see (2.7)) and zero, if the consequence does not hold (see (2.8)). Given the graph topology, the (posterior) probability of the hypothesis being

[7]Using terminology of Bayesian statistics: The better an indicator is, the smaller the error probabilities $q = P(Con|\overline{Hyp})$, $1 - p = P(\overline{Con}|Hyp)$.

true, is not only determined by the incoming evidence (*Rep*), but also by the reliability node *Rel* (Bovens and Hartmann 2003, p. 92, Equation 4.5). This probability can be computed directly from the conditional probabilities specified at the nodes in the Bayesian network.

Central to Bovens and Hartmann's results is their parameter a_i. Whenever $a_i = 0.5$, the instrument is a proper randomiser delivering a positive report with a 50/50 chance. However, whenever $a_i > 0.5$, then we are dealing with a "yes-man", whereas when $a_i < 0.5$ the instrument tends to be a "naysayer". An immediate consequence of this is that when $a_i < 0.5$, consistency of *positive* reports from the same instrument tends to disfavour the belief that this is indeed a randomiser, because you think it to be unlikely that a naysayer delivers two positive reports in a row. Since the only alternative to the instrument being a randomiser is that it is perfectly reliable, the confirmatory boost of two positive reports is higher when coming from the same instrument, than from two distinct ones; which leads to VET failure for $a_i < 0.5$. This dynamics occurs more or less invariably in the other scenarios which they analyse.

2.1.2 Bovens and Hartmann's Analysis of the Variety of Evidence Thesis

Bovens and Hartmann analyse three pairs of competing strategies, and compare the confirmatory value of one against the other, under various *ceteris paribus* conditions. See Figure 2, and Table 1.

	One Consequence	Two Consequences
One instrument and one measurement	A	/
One instrument and two measurements	B	C
Two instruments and two measurements	D	E

Table 1: The five possible experimental strategies as a function of how many consequences are tested, how many times, with how many instruments.

Bovens and Hartmann mathematically and graphically represent the fact that the consequence of the hypothesis is tested with the same instrument by having the reports sharing the same reliability node (instead of having independent ones as their parents). This is analogous to the representation of error independence in structural equation modelling, where error independence encodes the assumption that any latent variable possibly explaining co-variance between variables as their common cause is excluded.

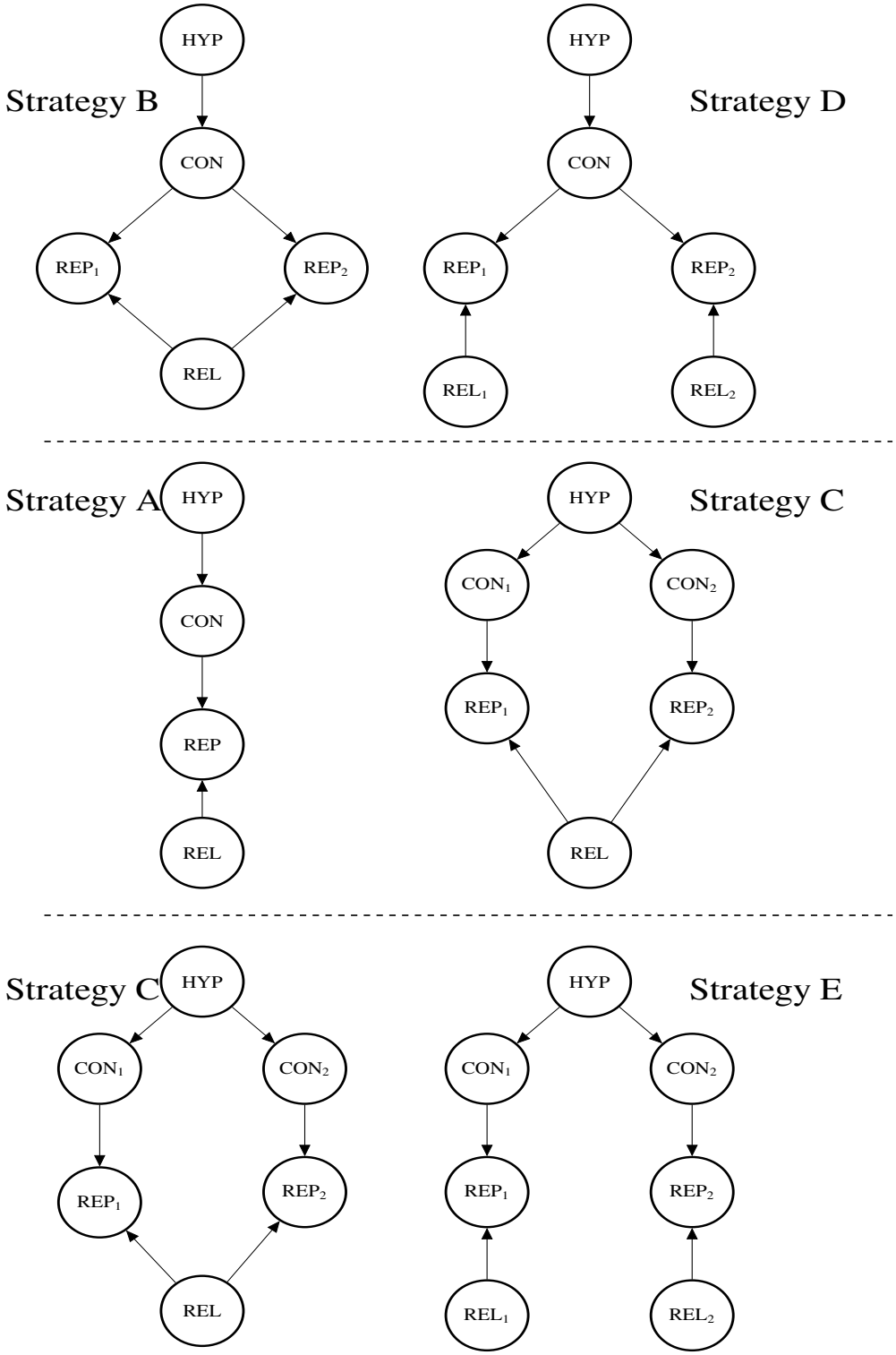

Figure 2: The three scenarios described in Bovens and Hartmann (2003): each row represents a scenario comparing two parallel strategies: B vs. D in the upper row, A vs. C in the middle, and C vs. E in the lowest row. *B vs. D* compares a situation where the scientist may collect two measurements concerning the same testable consequence of the hypothesis, from the same instrument (Strategy B) vs. the situation where the two measurements come from different instruments (D); *A vs. C* regards the situation where the scientist tests one consequence (A) vs. testing two consequences with the same instrument (C). *C vs. E* compares a situation where the scientist tests two consequences with the same instrument (C) with one where each consequence is tested by an independent instrument (E).

According to Bovens and Hartmann's conceptualisation of varied evidence, in each of such scenarios, the second body of evidence is more varied than the first body of evidence. This is because testing the consequence of the hypothesis with the same instrument, is a way to decrease variety. If variety of evidence is intended in this way, then the Variety of Evidence Thesis entails that, ceteris paribus, the posterior probability of the hypothesis of interest given two positive reports is larger under the second condition than under the first one for each of the three scenarios.[8]

Denoting by E the available evidence, by P the probability function for the less varied body of evidence, and P_1 the probability function for the more varied body of evidence, the VET can be stated as the requirement that

$$P(Hyp|E) < P_1(Hyp|E) \ . \tag{2.9}$$

Bovens and Hartmann provide explanatory rationales for the mathematical results of each scenario (Bovens and Hartmann 2003, pp. 94-107). We add in the following our considerations.

First Scenario

In the first scenario, that is, when the same consequence is tested twice with same vs. different instruments, the VET fails for values of ρ and a below .5, and we observe a trade-off between belief in the instrument being reliable and the rate at which it delivers positive reports, if it is a randomiser. The VET never fails when the instrument(s) are believed to be more reliable than not ($\rho > .5$), and in the remaining space it fails for low values of a, that is, it fails whenever one considers the randomising instrument to be a 'naysayer' (see explanation in previous section). We discovered curious dynamics for $N > 2$ (see also Bovens and Hartmann, unpublished). The

[8]We are not much interested in whether the comparison of posterior probabilities should be strict or not. We will say that the VET holds, if the second posterior is strictly larger than the first and say that the VET fails, if the first probability is greater or equal than the second probability. We think that this more adequately captures scientists' intuitions concerning the confirmatory support of varied evidence.

curves for more than 2 positive reports are particularly revealing: the higher the number of reports, the greater the area of VET failure, and this area increases in a. That is, the greater the number of reports, the greater a exists for which the VET fails: even when the probability of receiving positive reports from a randomising instrument is high ($a > .5$), the chance of receiving many of them from a reliable instrument that deterministically says yes if the hypothesis holds is still higher. Hence, our confidence in the instrument being reliable rather than a randomiser monotonically increases with increasing number of consistent reports.

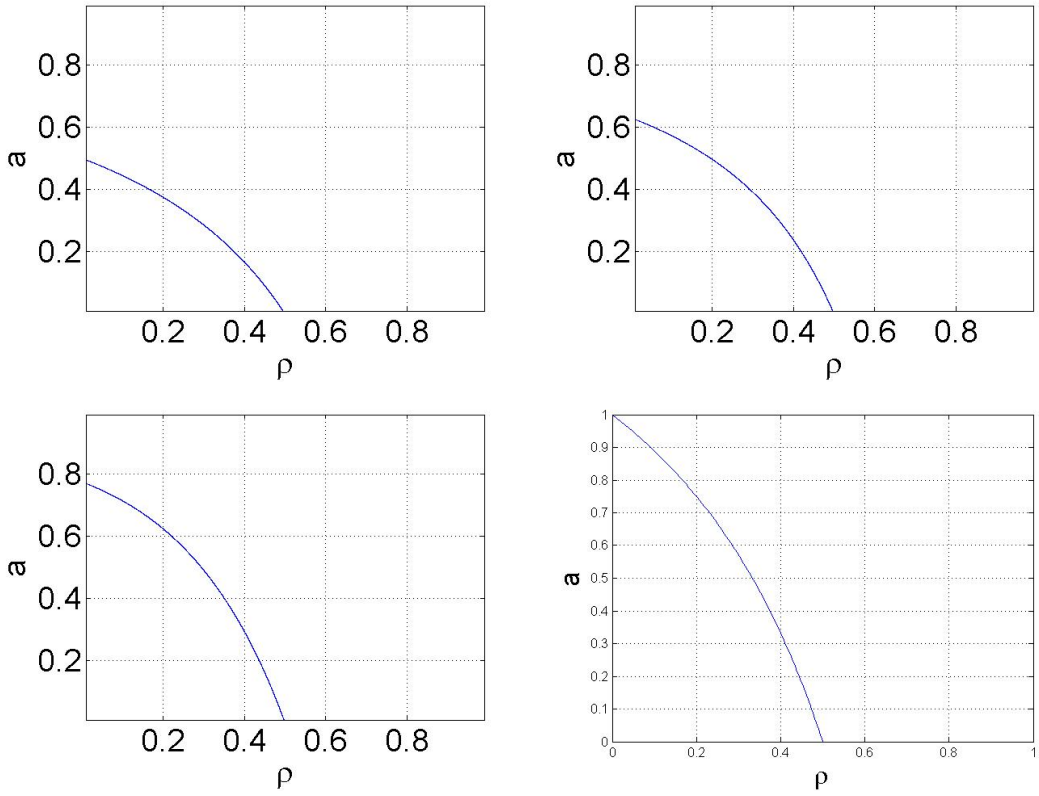

Figure 3: **Scenario 1 from (Bovens and Hartmann 2003, p. 97) for** 2 **positive reports (top left),** 4 **(top right),** 10 **(bottom left) and** ∞ **(bottom right)**. The VET fails in $\rho - a$-plane in the area under the curves. The plot on the lower right: for every point underneath the curve in the $\rho - a$-plane there exists a number N such that the VET fails for more than N items of evidence.

Second Scenario

The second scenario, that is, when the same consequence is tested once or twice with the same instrument, can barely be considered a case of contrast between varied vs. less varied evidence because in both settings evidence comes from one and the same instrument; what differs is the amount of reports received from it. Consequently, VET failure follows different patterns: we

test a consequence of the hypothesis[9], receive a positive report, and the posterior probability of the hypothesis increases. Then, for certain values of a, ρ, p, q, when we test a second consequence *with the same instrument* and we receive a second positive report, our confidence in the hypothesis *decreases*. In Figure 21 and Figure 22, we reproduce Bovens and Hartmann's results for the second scenario. Here, areas where the VET fails tend to be larger where i) the prior belief in the instrument being reliable is low (ρ is low), ii) the unreliable instrument is believed to be a yes-man (a is high) and iii) p and hence $q < p$ are small: the consequence is only loosely correlated to the hypothesis, and we do not expect to see many positive reports from a reliable instrument anyway, whether the hypothesis holds or not. As a consequence, receiving two positive reports about a consequence from the same instrument, instead of just one, tends to increase the belief that they are coming from an unreliable instrument rather than from a reliable one.[10]

Third Scenario

In the third scenario we go back to the epistemic dynamics already seen in the first one: the effect of two concurrent positive reports from a single instrument, rather than from two instruments, carries more confirmatory weight when both a and ρ are low (Figure 4). However here, as for the second scenario, also p and q play a role, but only when $a \approx 0.5$, that is, when the unreliable instrument is close to a proper randomiser. In this case, the VET fails when p and q are both high: you would expect a truth-tracking instrument to deliver positive reports most of the time, no matter whether the hypothesis holds or not, because the consequence would hold anyway. More often when the hypothesis is true – $p > q$ – but also when it is false – high q; hence receiving consistent positive reports from the same instrument reduces the belief in it

[9]We use scare quotes here, since also in the "varied" strategy, all reports come from the same instruments. Hence this is not so much a case of about varied vs. less varied evidence; but rather more vs. less evidence

[10]Think of a disease on which fever (the consequence of the disease) is supposed to fluctuate. Then if your thermometer always signals temperature, this makes you think that the thermometer is stuck. We thank an anonymous reviewer for this example. Further discussions can be found in Appendix C.

being a randomiser and increases the complementary belief in it being reliable.[11]

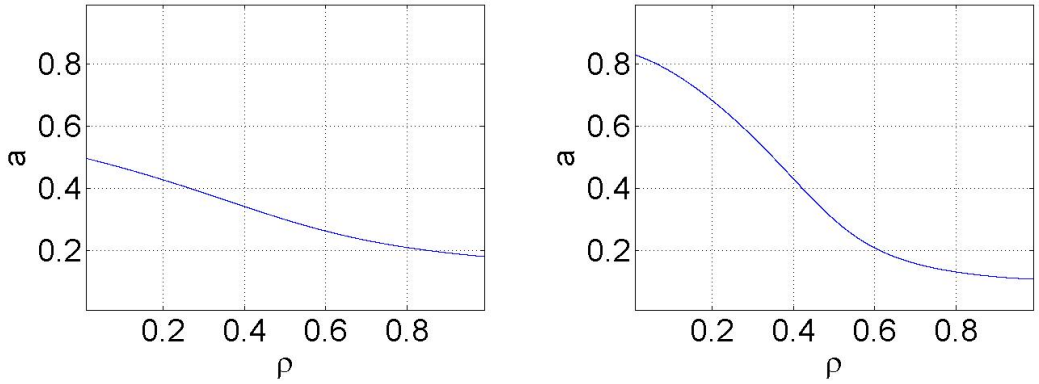

Figure 4: **Left: Figure 4.10 from (Bovens and Hartmann 2003, p. 102), Scenario 3.** The VET fails in the $\rho - a$-plane underneath the curve for $p = 0.9$ and $q = 0.1$. **Right:** Increasing the number of items of evidence to 50 while keeping $p = 0.9$ and $q = 0.1$ we can see that the area of VET failure increases. Interestingly, some parameter combinations of (ρ, a) where the VET fails for two reports do satisfy the VET for larger number of items of evidence.

2.1.3 The Variety of Evidence Thesis and Multiple Reports

Similarly to Scenario 1, we discovered that in the Bovens and Hartmann's model the areas of VET failure grow in all three scenarios for larger bodies of evidence. This is also something that contrasts with common intuitions about a growing body of evidence: one tends to think that the area of VET failure should not increase with an increasing number of reports: the more numerous the reports, the higher the epistemic import one would expect from receiving them from distinct sources. Bovens and Hartmann's model instead implies the opposite. This is because an increasing number of consistent positive reports from the same instruments monotonically decreases the probability that they come from a randomiser, whereas receiving

[11] Instead, p and q play only a marginal role when a is not close to 0.5, see Figure 15. For instance, for $a = 0.9$ the VET holds for almost the entire area of the relevant parameter space – that is, below the equality line $p = q$ – whereas it almost never holds for $a = 0.1$. When the unreliable instruments are believed to be yes-men, then one prefers to observe consistent positive reports from distinct ones: the VET holds. Instead, when they are believed to be naysayers, receiving consistent positive reports from one and the same instrument drops the belief in it being a randomiser, and hence the VET fails.

them from different instruments would not decrease such a probability to the same extent. See Figures 21–10 for plots and Appendix A.1 for the formal analysis.

This however contrasts with the common intuition that a long series of consistent positive reports may be due to bias rather than to the hypothesis being true ("too-good-to-be-true" evidence): in the real world, even reliable instruments might every now and then deliver false negatives. This is one reason why we moved to a different parametrisation of the model. The other one being consideration of random vs. systematic error in the picture. But first let's briefly present Claveau (2013), who explicitly takes systematic error into account in his contribution to the debate.

2.2 Claveau's Results

In Bovens and Hartmann's approach, all results depend on the reliability variable being strongly constrained as a binary variable (with two values: reliable and unreliable), and by delivering perfect information when it is reliable, and fully random information when unreliable. According to Claveau, this disregards the possibility of the instrument being systematically unreliable (biased).[12] This sort of unreliability does not root in randomness of reports, but rather in the instrument measuring something different than what it is intended to measure. Claveau makes the example of cases where an association between two variables is taken to be causal because a confounding factor creates a spurious correlation between them, and such correlation is systematically observed in repeated observations, unless one removes the confounder. These examples of unreliability regard the causal structure underpinning the statistical data. Other

[12] As also Claveau acknowledges, Bovens and Hartmann are aware of this limitation, see (Bovens and Hartmann 2003, Foonote 4, pp. 95-96): "Our model does not apply to unreliable instruments that do not randomize, but rather provide accurate measurements of other features than the features they are supposed to measure. In effect, our model exploits the coherence of the reports as an indicator that the reports are obtained from reliable rather than unreliable instruments. But if unreliable instruments accurately measure features other than the ones they are supposed to measure, then they will also provide coherent reports and so the coherence of the report is no longer an indicator that they were obtained from reliable instruments."

obvious cases of systematic error are biases in the experimental setting due to experimenters' (strategic) interference in the measurement process.

In particular, Claveau identifies two specific weaknesses in Bovens and Hartmann's argument: 1. reliability is modeled in such a way that fully dependent sources may still deliver different, independent, reports if they are unreliable; 2. They prove failure of the VET only for models where sources are either fully dependent or fully independent. Hence, Claveau goes on to propose a model where dependence comes in degrees, and where reliability becomes a ternary variable with values: fully reliable, positively biased, and negatively biased (no place is left for a randomising source). He then adopts a special measure of degree of independence δ and restates the VET as the requirement that the derivative of the posterior probability of the hypothesis over the interval from 0 to 1 of δ be non-negative:

> **Variety-of-evidence thesis.** Ceteris paribus, $\frac{\partial P^*_F(h)}{\partial \delta} > 0$ for all admissible values of ρ, α and δ. (Claveau 2013, p. 109)

The posterior probability of the hypothesis increases as we marginally increase the degree of independence of the evidential sources. By using this formalisation of the VET, and, in particular, δ as a measure of the degree of dependence of observations, Claveau shows that the parameter space where VET fails is smaller than in Bovens and Hartmann's analogous case (Claveau 2013, Footnote 10, p. 109-110).[13][14]

3 Our Model

The distinctive feature of Bovens and Hartmann's vs. Claveau's framework is that the former models unreliability in the form of a randomising instrument, whereas the latter models it

[13]Claveau's model is here not extended to multiple reports since our model is much closer to the Bovens and Hartmann approach and since it is not clear how to extend his notion of dependence to the multiple report case. Our mention of Claveau's model intends to acknowledge his incorporation of biased sources in a formal model of hypothesis confirmation.

[14]The model is extended to consequences which are also dependent to a degree in Claveau and Grenier (2019) and further discussed in Casini and Landes, Landes.

exclusively as systematic error. In both cases, the signals are fully disconnected from the state of nature: in (Bovens and Hartmann 2003) the source gives a positive or a negative signal at a given rate, no matter what; in Claveau (2013), the biased source always gives the same signal (positive or negative), no matter what. However, *neither of them incorporates any notion of random error*.

Hence, we propose a third model where reliable instruments are affected by random error (and therefore deliver imperfect information), and unreliable instruments are positively biased but not deterministically so.[15] Our model shares the variables and the network topology of Bovens and Hartmann's model. To formalise random error, we give up the deterministic relationships between (un-reliable) sources, state of the world and reports.[16] In our model the instrument is:

1. either reliable, but always associated with a certain amount of random error.

2. or it is positively biased with respect to the reliable instrument (results can be applied to negative bias mutatis mutandis).

The biased source is not deterministic, i.e., the probability of receiving a positive report is strictly less than one. This guarantees that signals coming from the biased source are still relevant, in that they are probabilistically relevant with respect to the investigated hypothesis. They are not disconnected from reality.

In our model the agent is confronted with the hypothesis that the source is either imperfectly reliable but somewhat *fair with respect to some domain-related standards*, or that data are being distorted towards accruing positive evidence.

In this setting, one would expect VET to invariably hold, since independent sources of evidence would help decrease the probability that consistency of reports is due to bias, and thereby jointly confirm the hypothesis more than if they came from the same source. Yet, also in our model VET fails for specific combinations of the model parameters.

[15] In the following it will be only positively biased; exploration of different combinations will have to wait for future studies.

[16] Technically, we allow for the possibility that an unreliable instrument is a complete randomiser; although this is a borderline case in the continuous space of possibilities.

3.1 Model Parameters

Our model differs in the following from its predecessors:

1. Reliable instruments are not fully reliable, and may deliver false positives (or false negatives), although with low probabilities $\epsilon_+, \epsilon_- > 0$.

2. Unreliable instruments are positively biased with respect to the reliable ones, with probability γ of delivering a positive report, when the consequence does not hold (false positive), and probability α of delivering a positive report, when the consequence holds (true positive).

Positively biased instruments may represent e.g. the case of "selective reporting" under information asymmetry analysed in the economic literature of games with private information (see Osimani et al. (2020) for a recent review).[17] Positively biased sources deliver true positive results at higher rate (α) than $1 - \epsilon_+$ (which is the true positive rate for reliable sources), and false positive results also at a higher rate (γ) than a reliable source does (ϵ_-).[18]

The novel parameters are hence α, γ and ϵ_+, ϵ_-. Among them we have the following relations: $\alpha > 1 - \epsilon_+$ and $\gamma > \epsilon_-$. These inequalities model positive bias in relative terms. It is natural to require that $\alpha > \gamma$, which says that, if positively biased, the probability of the instrument delivering positive reports should be higher when the consequence holds, rather when it does not. *This assumption is what leaves some room for the biased instrument to still deliver*

[17] We consider that the probability distribution over the hypotheses space regarding whether the instrument is of type 1. or 2. (that is, reliable, or positively biased) is such that, the hypothesis of the instrument being positively biased and negatively biased exclude each other. That is, as soon as one considers that the source might be affected by some sort of systematic error (due to confounders or bias, or both), then this is estimated to bias results in one direction only, on the basis of one's expectations regarding the incentive structure of the testing system Osimani et al. (2020)).

[18] Strictly speaking, all these parameters need to be indexed by the particular report variable under consideration. To simplify the exposition and in anticipation of the ceteris paribus conditions (see below) these indexes are suppressed in the notation.

relevant information.Otherwise for $\alpha = \gamma$, the unreliable instrument delivers positive results independently of the truth of the consequence; in other words we have a randomising instrument in the sense of Bovens and Hartmann (see also Section 3.7).[19]

On the other hand, a positively biased instrument tends to deliver false negative results at a lower rate than a reliable instrument does: $1 - \alpha < \epsilon_+$. Hence, a biased source will tend more often than the reliable one to claim that the hypothesis is true when it is; its true positives rate is higher than that of the reliable source.

Finally, we assume non-extreme conditional probabilities, all probabilities are strictly between zero and one.

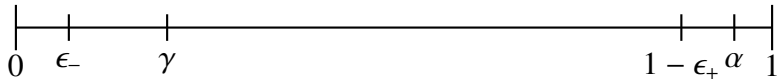

Figure 5: Parameter configuration for two alternative hypotheses regarding the reliability of the instrument: when the instrument is *Rel* it delivers true positives with rate $1 - \epsilon_+$ and false positives with rate ϵ_-. When the instrument is $\overline{Rel}$, that is, positively biased, it delivers true positives with rate α and false positives with rate γ. Our constraints on our model are such that $\epsilon_- < \gamma$ and $1 - \epsilon_+ < \alpha$. If it were the case that $\gamma < \epsilon_-$ and $\alpha < 1 - \epsilon_+$, we would consider a negatively biased instrument. $0 < \gamma = \alpha < 1$ is a "randomising" instrument in the Bovens and Hartmann sense.

While we are here only interested in positively biased instruments 1, there are further parameter configurations which correspond to different types of instruments:

1. $\epsilon_- < \gamma$ and $1 - \epsilon_+ < \alpha$ (positively biased),

2. $\gamma < \epsilon_-$ and $\alpha < 1 - \epsilon_+$ (negatively biased),

3. $\gamma < \epsilon_-$ and $1 - \epsilon_+ < \alpha$ and

4. $\epsilon_- < \gamma$ and $\alpha < 1 - \epsilon_+$.

[19]Conceptually, a directed (here positively directed) bias means that an instrument is more likely to deliver a result in direction of the bias than a reliable unbiased instrument. Graphically speaking, γ and α are shifted to the "right of" ϵ_- and $1 - \epsilon_+$, respectively; see Figure 5. For a negatively biased instrument, matters would simply reverse: α and γ would be closer to zero than $1 - \epsilon_+$ and ϵ_-, respectively.

3 and 4 model two different instruments which differ in terms of their respective accuracy only. In both cases the more accurate instrument has the lower probability for false positives and the higher for true positives. However, bias in one direction or the other is absent.

Variable	Intended Interpretation	(Conditional) Probabilities	Model
HYP	Hypothesis of Interest	$0 < P(Hyp) < 1$	B&H, we
CON_i	Testable Consequence	$0 < P(Con\|\neg Hyp) < P(Con\|Hyp) < 1$	B&H, we
REL_i	Reliability of Instrument	$0 < P(Rel) = \rho < 1$	B&H, we
REP_i	Report	see below	B&H, we
Rel_i	Reliable Instrument	$P(Rep\|Con, Rel) = 1$	B&H
		$P(\neg Rep\|\neg Con, Rel) = 1$	
$\overline{Rel_i}$	Unreliable Instrument	$0 < P(Rep\|Con, \neg Rel) = a < 1$	
		$0 < P(Rep\|\neg Con, \neg Rel) = a < 1$	
Rel_i	Reliable Instrument	$0 < P(Rep\|Con, Rel) = 1 - \epsilon_+ < 1$	we
		$0 < P(Rep\|\neg Con, Rel) = \epsilon_- < 1$	
$\overline{Rel_i}$	Unreliable Instrument	$1 > P(Rep\|Con, \neg Rel) = \alpha > 1 - \epsilon_+$	
		$1 > P(Rep\|\neg Con, \neg Rel) = \gamma > \epsilon_-$	

Table 2: Overview of employed variables, their intended interpretation and (conditional) probabilities in Bovens and Hartmann's and our model. To increase readability, we use ¬ to denote negation in this table.

We understand the VET here the same way as Bovens and Hartmann, and hence assume that the body of evidence which is more varied in the sense expressed by them, should enhance the posterior probability of the hypothesis more strongly than a less varied body of evidence, that is, (2.9) ought to hold. We shall consider the three scenarios presented by Bovens and Hartmann.[20]

3.2 Belief Dynamics

Although our instruments are closer to real ones, the belief dynamics are ideal in the sense that the agent is assumed to be uncertain about whether the testing instrument(s) are positively

[20]To do so, we make the usual ceteris paribus assumptions that all reliability variables REL have the same prior probabilities and that all report variables REP have the same conditional probabilities. The variables HYP and CON are the same under both conditions and hence receive equal (conditional) probabilities.

biased or not, but then to exactly know the rate of true and false positives in each case: $1-\epsilon_+$ and ϵ_- respectively, if the instrument is reliable, and α and γ, respectively, if it is biased. This is due to the reliability variable being binary. Anyway, considering non-binary reliability variables – for example adding values for more types of unreliability – does not pose any conceptual problem (Olsson 2005, Section 4.3). It merely requires the specification of further priors in those types of unreliability and the corresponding conditional probabilities of reports.

3.3 Scenario 1

We show in Appendix B that for two positive reports the VET does fail:

Theorem 1. *For all* $p \in (0,1)$, $q \in (0,p)$, $\rho \in (0,1)$, $\epsilon_+, \epsilon_- \in (0,1)$, $\alpha \in (1-\epsilon_+, 1)$ *and* $\gamma \in (\epsilon_-, 1)$ *the VET fails, if and only if*

$$0 < \gamma_2 \leq \gamma \leq \gamma_1 < 1 \ ,$$

where γ_1 *and* γ_2 *are the following parameters*

$$\gamma_2 := \frac{\epsilon_- \cdot [2\rho(1-\epsilon_+) + \alpha(1-2\rho)]}{(2\rho-1)(1-\epsilon_+) + 2\alpha(1-\rho)} < \frac{\epsilon_-}{1-\epsilon_+} \cdot \alpha =: \gamma_1 \ .$$

That $P(Hyp|E) - P_1(Hyp|E) = 0$, if and only if $\gamma \in \{\gamma_1, \gamma_2\}$ is established in Proposition 8. γ_1 is independent of ρ. Since $\gamma_2 < \epsilon_-$ (Proposition 11), it follows that $\gamma = \alpha\epsilon_-/(1-\epsilon_+)$ is a necessary and sufficient condition for the posteriors of the hypothesis to be equal (within the relevant parameter space).

Therefore, the VET fails whenever:

$$\frac{\gamma}{\alpha} < \frac{\epsilon_-}{1-\epsilon_+} \ . \tag{3.1}$$

That is, when the ratio of false to true positives is lower for the biased instrument than for the one which is affected by random error only (proofs are in the appendix). In case the ratio of false to true positives is lower for the positively biased instrument than for the reliable one, a sequence of two or more positive reports *coming from the same instrument* boosts both the

belief in the instrument being positively biased and, because this also has a more favourable ratio of true vs. false positives, in such positives being true ones, more than it would happen if the results came from different instruments. The VET fails.

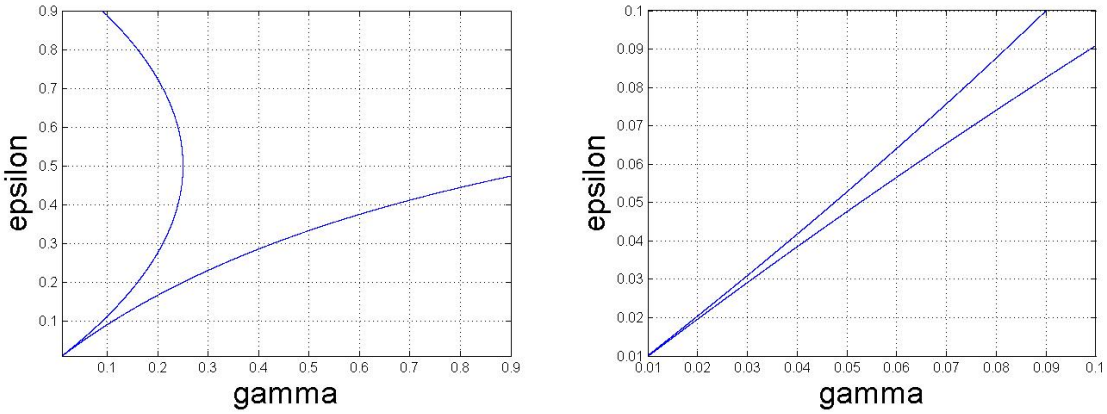

Figure 6: **Scenario 1: the $\gamma - \epsilon$-plane for N = 2**, the curves along which posterior degrees in the hypothesis are equal for $\alpha = 1$. ($\alpha = 1$ was chosen to ensure that all $\epsilon > 0$ are permissible values.) The VET holds for the points under the lower curve and on the left of the upper curve. Points on the left of the upper curve indicate that the VET holds in this area – which is irrelevant for our current purposes – since $\gamma > \epsilon$ holds there. The right graph focuses on the area where the random error ϵ is less than 10%. The VET fails in the narrow region between the diagonal, $\gamma = \epsilon$ (not pictured), and the lower curve ($\gamma = \epsilon/(1 - \epsilon)$).

An analytic narrative for these results can be provided as follows:

1. Given the model parameters (see Fig. 5), a positive report is more likely if an instrument is systematically biased than if it is affected by random error only;

2. the probability that the instrument is systematically biased hence increases as the agent receives positive reports from this instrument: $P(\overline{Rel}|Rep) > P(\overline{Rel})$;

3. given the model topology (see Fig. 6, e.g., the contrast between Strategies B and D), the second positive report in the more diverse strategy is taken into account based on the prior probability $P(\overline{Rel})$. In contrast, the second positive report in the less diverse strategy (e.g., strategy B) is taken into account based on the posterior probability $P(\overline{Rel}|Rep)$;

4. given points 2 and 3, the agent will judge it more likely that the second positive report comes from a systematically biased instrument if she uses a less diverse strategy than if she uses a more diverse strategy;

5. hence, if a positive report is more confirmatory when it comes from a systematically biased instrument than when it comes from a non-biased instrument (because positives are more likely to be true in the former case), then multiple positive reports from a less diverse strategy will be more confirmatory than multiple reports from a more diverse strategy;

6. the counter-intuitive antecedent in the point 5 holds, iff $\frac{\gamma}{\alpha} < \frac{\epsilon_-}{1-\epsilon_+}$ holds;

7. our mathematical result is thusly epistemically explained.

In our model as well as in Bovens and Hartmann's and Claveau's model, VET failure follows from non-varied evidence providing "higher order" evidence about the source of the reports.

All of our results also hold, if $\epsilon_- \geq 1 - \epsilon_+$. This dovetails nicely with (Landes 2020, Sections 6.1 and 6.2) where the fate of the VET also does not depend whether evidence has a Bayes factor greater that one.

By plotting the difference of the posteriors in the two conditions for our first scenario on the $\gamma - \epsilon$-plane we obtain Figure 6.[21] We observe three patterns:

1. as γ and ϵ tend to 0 the area of VET failure diminishes and becomes eventually negligible.

2. for $\gamma \gg \epsilon$ the VET holds;

3. for high γ and ϵ the VET fails as $\epsilon > \gamma$;

1) For very low values of γ and ϵ, the instruments are believed to be precise (if reliable) or with very low bias (if they are biased): independent items of evidence are more confirmatory than same ones coming from an identical source: the area where the VET fails becomes negligible (see Figure 6).

[21] To cut down the number of parameters for our graphical explorations, we assume that errors of the first and second type are equally likely, $P(\overline{Rep}|Con, Rel) = \epsilon_+ = \epsilon_- = P(Rep|\overline{Con}, Rel) =: \epsilon$. In Figure 6 α is set to 1. However, things do not change by varying this parameter, to the extent that it remains in the permissible range of our model, that is: α and γ are shifted to the right of ϵ ans $1 - \epsilon$, respectively.

2) When evidence from the biased instrument is highly biased relatively speaking (γ much larger than ϵ), we have a high suspicion of positive results being false when coming from the same instrument: a consistent series of positive reports is more confirmatory when coming from independent sources and confirmation is greater for varied evidence. The VET holds.

3) For high values of γ *and* ϵ both instruments tend to be sloppy and unreliable; the VET fails whenever the ratio of false to true positives is lower for the biased instrument, because two consistent reports from the same instrument increases the probability that they are true positives, *via increasing the probability that they come from the more accurate instrument between the two.*

3.4 Scenario 2

We recall that the second scenario compares a case where only one report is received, to a case where multiple reports are received from the same instrument, and "VET failure" is conceived as the hypothesis receiving less confirmatory support from multiple reports than from only one. We only found VET failures for extremely strong biases, $\gamma \approx \alpha$.[22] To illustrate our point we show Figure 18. We note that the number of reports matters only to a relatively small degree. This holds true for even larger N (figures not shown here).

The area of VET failure is negligible (both in comparison to Bovens and Hartmann and in absolute terms). More and more reports coming from the same instrument tend to confirm more than just one report, unless bias is extremely large, and hence evidence tends to confirm that the observed positives are likely to be false ones. [23]

In Bovens and Hartmann's model, the area of "VET failure" is larger for higher values of a

[22]Hence small ϵ: since in our model $\epsilon < \gamma$ and $1 - \epsilon < \alpha$ the area of VET failure admissible for our model is limited to the space where both inequalities hold.

[23]The equation determining the fate of the VET, see (B.4), is polynomial in the parameters $\alpha, \rho, p, q, \epsilon, \gamma$. Even for just two reports ($N = 2$ and $\epsilon_+ = \epsilon_- = \epsilon$) the polynomial is much too large to be tractably solved. To investigate further, we varied the relevant parameters, including the number of reports, N. The role of ρ is only marginal and nonlinear. For $N = 2$ the area of "VET failure" tends to slightly increase with increasing ρ up to $\rho = 0.5$ and then to diminish again, whereas it increases monotonically with increasing ρ for $N = 10$, see

and smaller values of ρ, that is when the randomiser is believed to be a yes-man and there is low confidence in the instrument being reliable. Analogously, in our model, the VET "fails" only for extremely high values of γ, approximating α: receiving numerous consistent reports from the same instrument increases the belief that we are dealing with a yes-man (since here $\gamma \approx \alpha$, the biased instrument tends to deliver irrelevant signals).

3.5 Scenario 3

In the third Scenario, the curve $\gamma = \alpha\epsilon_-/(1-\epsilon_+)$ – and, therefore, the ratios γ/α and $\epsilon_-/(1-\epsilon_+)$ – is again playing a key role:

Theorem 2. *For all* $p \in (0,1)$, $q \in (0,p)$, $\rho \in (0,1)$, $\epsilon_+, \epsilon_- \in (0,1)$, $\alpha \in (\epsilon_+, 1)$, $\gamma \in (\epsilon_-, 1)$ *and all* $N \geq 2$, *if* $\gamma = \alpha \cdot \frac{\epsilon_-}{1-\epsilon_+}$, *then*

$$P(Hyp|E) = P_1(Hyp|E) \ .$$

However, as we can see from Figure 7 a second curve (in green), along which $P(Hyp|E) = P_1(Hyp|E)$ holds too, intersects the curve $\gamma = \alpha\epsilon_-/(1-\epsilon_+)$.[24] This curve is outside the specified parameter bounds though, in that $\alpha < 1 - \epsilon_+$ holds there.

Our graphical exploration of the six-dimensional parameter space $(p, q, \rho, \alpha, \epsilon, \gamma)$ for larger N, did not reveal any difference with respect to the two-reports case. We saw for all parameter values we checked that the curve $\gamma = \alpha\epsilon/(1-\epsilon)$ continued to be the only area where both posteriors were equal. Coherently with all results, that is, that VET failure depends on the relative preponderance of false vs. true positives for the biased vs. unbiased instrument, in the $\epsilon - \gamma$ plane, VET holds for larger γ and fails for smaller γ (Figure 19 and Figure 20 in the Appendix illustrate this for the subspace with $\epsilon, \gamma \in (0, 0.2)$).

As in Scenario 1, the status of the VET only depends on whether γ/α is greater or less than $\epsilon/(1-\epsilon)$ within the considered parameter values, independently of N, ρ, p, q. This is so for

Figure 18.

[24] This second curve is found by inserting the parameter values $(N = 2, \rho = 0.25, q = 0.75, p = 0.9, \alpha = 0.975)$ and solving the polynomial in the variables γ, ϵ symbolically in Octave

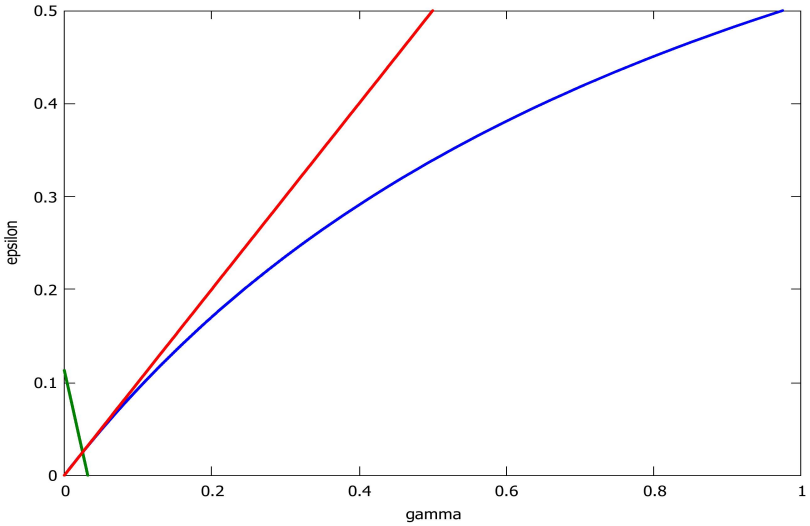

Figure 7: **Scenario 3: The $\gamma - \epsilon$-plane**. The relevant parameter space $\gamma > \epsilon$ (below the red curve) is divided in two parts by the curve $\gamma = \alpha\epsilon/(1-\epsilon)$ blue curve and the curve $\gamma = -5\epsilon/3 + \sqrt{11680\epsilon^2 - 213360\epsilon + 226161}/120 - 77/20$ green curve. The VET holds for values below the blue and on the right of the green curve. Unlike in Scenario 1, the VET now also fails for tiny random errors ϵ. However, note that $\alpha = 0.975$ and hence $\epsilon > 0.025$ holds in our model. This means that the small triangle near the origin where VET fails (bordered by the green and red curve as well as the x-axis) is *outside* the allowed parameter values of our model.

analogous reasons to Scenario 1.

3.6 Multiple Reports

The fate of the VET is robust to multiple reports in our model. That is, the area of VET failure in the phase space tends not to increase or decrease with increasing N.

Scenario 1: Our graphical exploration of the parameter space for larger N indicates that the VET fails for all $\gamma < \alpha\epsilon/(1-\epsilon)$ and all N. The cases of $N = 5$ and $N = 50$ consistent positive reports are shown in Figure 16 and Figure 17.[25] For all $N \geq 3$, the VET continues to hold for $\gamma > \gamma_1$ and continuous to fail for $\gamma \leq \gamma_1$ that are also close to γ_1 (Proposition 10).

The fate of the VET is also relatively robust under changes of N in Scenario 2 and Scenario 3 (Sections 3.4 and 3.5).

[25] In Figures 16 and 17, we also notice that the area of VET failure increases for $\epsilon > \gamma$ for increasing N which also depends on ρ. The dependence on ρ does not come as a surprise given the explicit dependence of γ_2 on ρ in Theorem 1. However, $\epsilon > \gamma$ does not correspond to a biased instrument and is hence outside our model, see Section 3.7 for discussion.

This shows that VET failure in our model does not depend much on the amount of evidence, but rather on the dependency structure among items of evidence and the ratio of false-to-true positives associated with the testing instruments.

3.7 Bovens and Hartmann's Model as a Limiting Case

Setting $\epsilon = \epsilon_+ = \epsilon_- = 0$ and $P(Rep|Con, \overline{Rel}) = \alpha = P(Rep|\overline{Con}, \overline{Rel}) = \gamma = a$ we recapture the Bovens and Hartmann model.[26] Their model does not allow γ and α to be teased apart. In their approach, $P(Rep|\overline{Con}, \overline{Rel})/P(Rep|Con, \overline{Rel}) = a/a = 1$ is much different from $P(Rep|\overline{Con}, Rel)/P(Rep|Con, Rel) = 0/1 = 0$. So, epistemic dynamics and VET failures revolving around $\gamma/\alpha = \epsilon_-/(1 - \epsilon_+)$ are inconceivable from within the Bovens and Hartmann model, see Fig. 8 for an illustration.

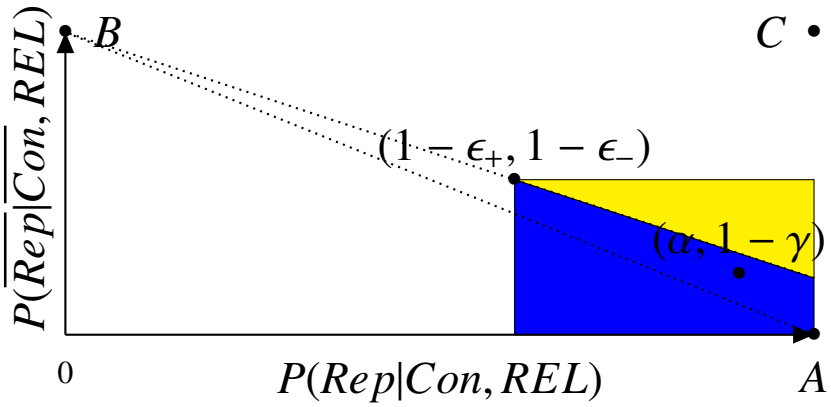

Figure 8: Plane of probabilities of true positives and true negatives. Reliable instruments in Bovens and Hartmann's and Claveau's sense are perfect (C). Unreliable instruments in Bovens and Hartmann's sense are represented by one point $(a, 1 - a)$ on the dotted line connecting A and B. Unreliable instruments in Claveau's sense are either at A (positively biased) or at B (negatively biased). Our unreliable instruments are represented by one point on the right and below the reliable ones (yellow and blue area). $\alpha/\gamma > (1-\epsilon_+)/\epsilon_-$ holds in the yellow area (VET fails). The opposite is true in the blue area (VET holds).

Indeed, for parameter choices approximating the Bovens and Hartmann model, $\epsilon \approx 0$ and $\gamma \approx \alpha$, the posterior probability of the hypothesis in our model will approximate the posterior in the Bovens and Hartmann model.[27] For such parameter values, the results obtained by Bovens

[26] This violates our assumptions of non-extreme probabilities ($\epsilon_-, \epsilon_+ > 0$) and random error ($\alpha > 1 - \epsilon_+$).

[27] This follows from the continuous dependence of the posterior probability distribution on

and Hartmann regarding the status of the VET (in all three scenarios) tend to hold in our model, too.

Hence, these cases allow us to explain some VET failures as being 'artefacts' of Bovens and Hartmann's model. In Scenario 2, we only found VET failures for $\gamma \approx \alpha$ and small ϵ, see Figure 18. These are precisely the limiting conditions under which our model converges to their model.

In our Scenario 3, we saw in Figure 7 that there is a small area near the origin, where the VET fails for $\gamma > \epsilon$. In this area, if $\alpha < 1 - \epsilon$, both the false positive (γ) and the true positive rates (α) of the unreliable instrument get closer to each other, that is, we are in case 4 in our list of possible models. This means that the reliable instrument is more precise than the unreliable one, whereas the unreliable one is more noisy. Plotting VET failures in the $p-q$-plane, we find in Figure 9 for $\alpha < 1 - \epsilon$ that our model is reminiscent of Figure 10 which depicts the VET failure in Scenario 3 in the Bovens and Hartmann model. The area of VET failure significantly increases with decreased α (from 0.995 to 0.985 and then 0.975).

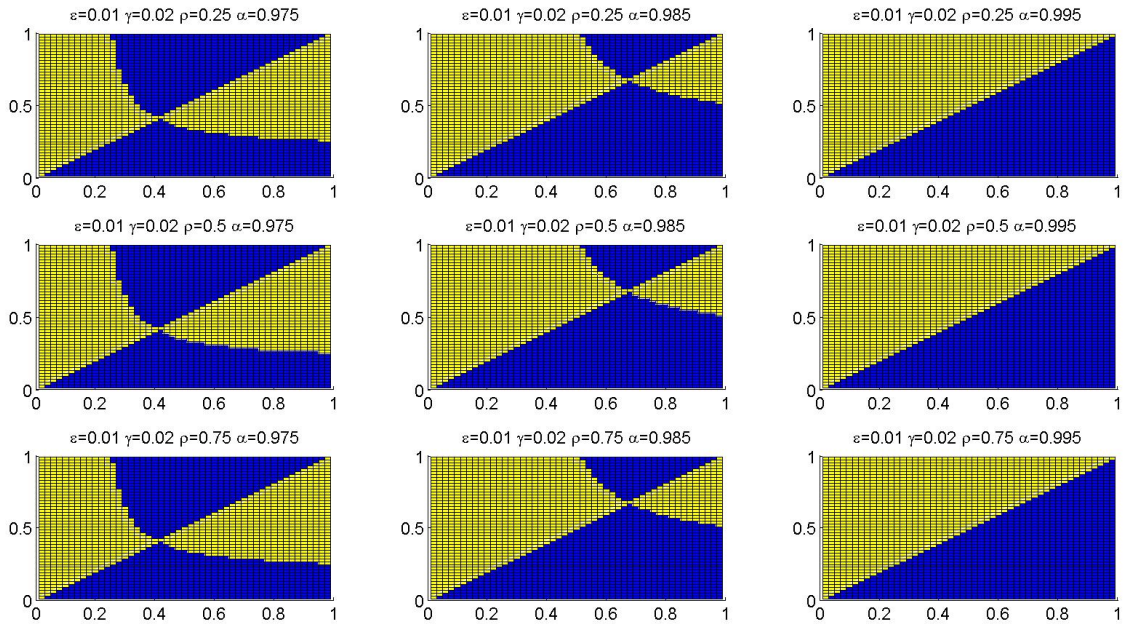

Figure 9: **Scenario 3 in our model: The p − q-plane for fixed $\gamma = 0.02, \epsilon = 0.01, \mathbf{N} = 2$ and varying ρ, α.** Yellow indicates the area of VET failure. Within a column of a set ρ varies: top $\rho = 0.25$, middle $\rho = 0.5$, bottom $\rho = 0.75$. Within a row α varies: left $\alpha = 0.975$, middle $\alpha = 0.985$, right $\alpha = 0.995$.

the input parameters (conditional probabilities).

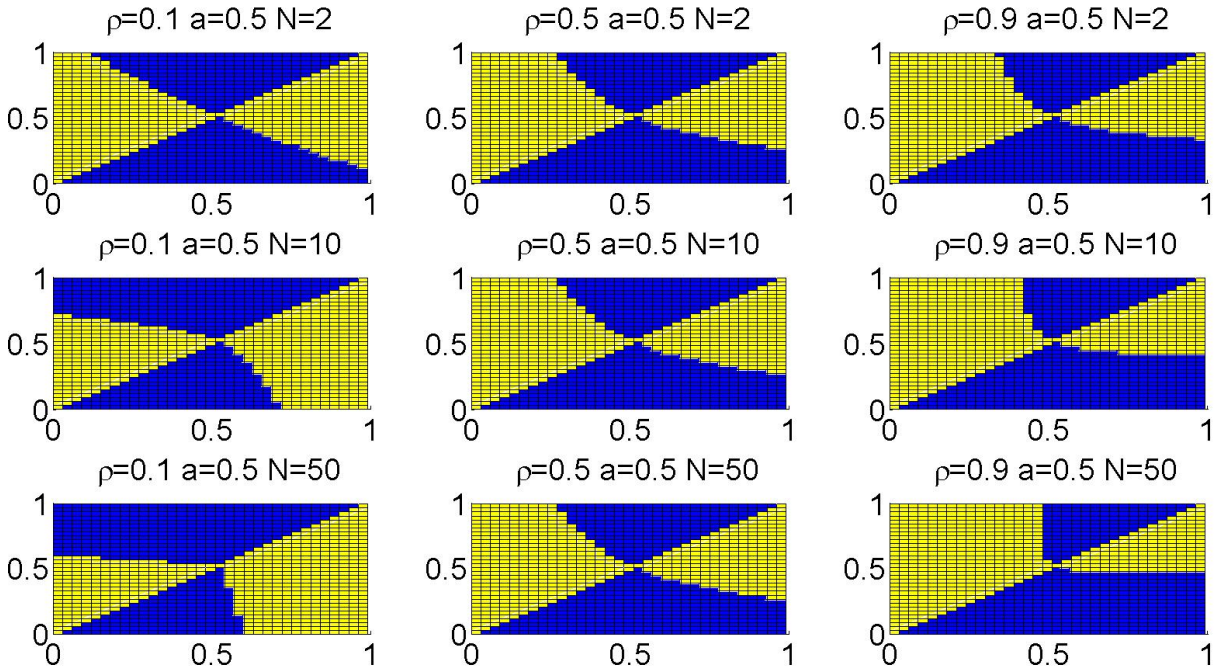

Figure 10: **Scenario 3 in Bovens and Hartmann model: The p, q-plane for varying ρ, N** and fixed **a = 0.5**. Yellow indicates the area of VET failure, blue means that the VET holds. Within a column N varies: top $N = 2$, middle $N = 10$, bottom $N = 50$. Within a row ρ varies: left $\rho = 0.1$, middle $\rho = 0.5$, right $\rho = 0.9$. When $a = 0.5$ (or close to this value), then the instrument is a randomiser par excellence: it delivers positive and negative reports half of the time, no matter what the truth is. In this case, the VET tends to fail for q and $p \geq .5$, that is, for consequences which have high rate of both true and false positives: one expects consistent positive reports from reliable instruments no matter whether the consequence holds or not; hence, if the series comes from a single instrument, this boosts the degree of probability that the instrument is reliable and, since the indicator is anyway positively relevant ($p > q$), this does more confirmatory work with respect to the hypothesis than having the same series from different instruments.

4 Conclusion and Outlook

With the present paper we aimed to: a) contribute to the debate on the Variety of Evidence Thesis by testing the results obtained in previous work on a different model of scientific inference, in that we add both random and systematic error to the picture, and b) investigate the epistemic dynamics that develop under such conditions. With respect to these two aims:

1. We verified VET failure for a model which maintains the topological characteristics of Bovens and Hartmann's but differs from it – and from Claveau's one – by incorporating both random error and undeterministic bias, as possible characteristics of the instrument, and by eliminating randomisers. We show that VET holds for a much greater area of the

phase space.

2. We spelled out the theoretical grounds and implications for the curve ($\gamma = \alpha\epsilon_-/(1 - \epsilon_+)$) along which posterior probabilities are equal, that is, more vs. less varied evidence conditions have the same confirmatory boost.

3. We found VET failure in our model to be robust with respect to the number of positive consistent reports obtained, in contrast to Bovens and Hartmann's model.

4. We found that our approach mathematically encompasses Bovens and Hartmann's approach. We hence were able to explain VET failures in our second scenario as artefacts of their approach.

Our analysis of the models brings to light how distinctive ways to model (un)reliability impact on the inferential import of consistent results. Bovens and Hartmanns have the counter-intuitive results that the area for which the VET fails grows with increasing number of reports from the same instrument. This contrasts with the intuition that "too much" consistency may speak for bias rather than truth, since in science one works under the assumption that a certain dose of inaccuracy – random error – inevitably affects any measurement instrument. Our model (see Section 3) tracks such reasoning in that the probability of true positives for a reliable instrument affected by random error is lower than the true positive rate for a positively biased instrument. Hence, in our setting, a series of consistent reports from the same instrument, does not necessarily increase the belief that the instrument is reliable, as it happens in Bovens and Hartmann's case, via decreasing the probability that it is a randomizer, but rather boosts the belief that it is biased. Therefore our model pays heed to the "too-good-to-be-true" evidence intuition. However, in our case, the VET fails exactly via exploiting awareness that the biased instrument is also less noisy than the fairer one.

Our analysis connects to the emphasis on "power" as a remedy against the so called "reproducibility crisis" Collaboration (2015), Begley and Ellis (2012), Prinz et al. (2011), Osimani et al. (2020). The reproducibility crisis relates to repeated failures of replication studies to reproduce the results delivered by original ones Ioannidis (2005). Most students agree that this phenomenon is originated by the reliance on low-powered studies, whose positives are likely

to be false Button et al. (2013), Lakens and Evers (2014), Smaldino and McElreath (2016): low-powered studies tend to work as "randomisers" in Bovens and Hartmann's sense. However, selection of statistically significant and positive findings through various forms of biases, tend to produce positively biased evidence and filter out negative or insignificant findings. Thus published studies are the result of complex data-generating processes affected by various combinations of random and systematic error influencing the output at various stages of the process. The final outcome may be a true or a false positive with varying probability depending on this process. The general doctor of our first vignette, as well as the other protagonists of the following ones, are exactly confronted with a dilemma regarding the data generating process behind the results of the studies. Our model helps to capture the interaction of beliefs about the evidence source and other dimensions of evidence, while identifying some essential dimensions of the inference leading from consistent evidence to hypothesis confirmation.

Our results rely on important assumptions though. In all models presented here the confirmatory boost of coherent evidence through (in)dependent evidence and instrument reliability are "incarnated" by a specific topological structure relating the evidence reports to the hypothesis. The structure itself and how it relates to scientific uncertainty has not been justified. The paper does not explicitly address the role that the Markov blanket plays in these models (see Wheeler and Scheines (2013)). In this respect, uncertainty regarding the reliability of the data generating process could invest not only the measuring instrument, but precisely the causal structure relating hypothesis and evidential reports more generally. Hence, uncertainty would not regard whether the measuring instrument is "reliable" or not, but whether we are dealing with a specific causal structure or another one.

The results that hold for the canonical case of a common cause model in which conditional independence is satisfied may not be a good approximation of what one would find if the conditional independence condition was "almost satisfied" — that is, where there is some small epsilon of association among the reports that is left after conditioning on the common cause. More generally, items of evidence may be related to one another and to the hypotheses in a number of different ways, and it is this structure that contributes essentially to whether coherence is confirmatory boosting or not. Wheeler and Scheines address this question, and show

that the relationship between associated (or not) witnesses reports and confirmation is mediated by the causal structure. Unlike the previous literature and unlike our study, they are precisely interested in the epistemic effects on hypothesis confirmation through probability propagation in such diverse structures. They study and compare scenarios where items of evidence for criminal settings display diverse conditional dependence relations among themselves and with respect to the hypothesis; e.g. cases where some of the items of evidence are parent nodes of the hypothesis of interest itself (for instance antecedents for motives). They also compare the performance of a number of measures of confirmation in keeping track of such epistemic dynamics, while we and the previous literature on the VET only consider the difference between prior and posterior probability of the hypothesis as our measure of confirmation. Anyway, they also identify cases where VET is violated, these are cases in which items of evidence that are incoherent among themselves confirm the hypothesis more strongly than a coherent "equally positive evidence set" (Wheeler and Scheines 2013, Proposition 6).

Their results suggest that a purely numeric analysis, such as this paper offers here, must be aided by explicit structural independence assumptions of the model, along with an accounting of the effect these structural assumptions have on the robustness of the model. This sort of consideration leaks also from the different role that the ρ parameter plays in Bovens and Hartmann's model and in ours: whereas in the former, this plays a prominent role, in our case it all but disappears in the results. This is exactly due to the different causal structure between evidence and hypothesis assumed in their model for the randomising instrument: as soon as the instrument is a randomiser ($REL = \overline{Rel}$), the connection between evidence and hypothesis is severed.

This does not happen for our unreliable instruments. As a consequence, in the hypothesis that the evidence is coming from a randomising instrument, the arrow from CON to REP is de facto deleted.

Hence, uncertainty as to whether the instrument we are dealing with is reliable, biased, or a randomiser may be modelled exactly via various topological structures for the epistemic dynamics representing scientific settings and learning scenarios more generally. This issue certainly deserves a separate future paper.

Funding

Barbara Osimani gratefully acknowledges funding from the European Research Council ('Phil-Pharm' grant 639276)

Jürgen Landes gratefully acknowledges funding from the European Research Council ('Phil-Pharm' grant 639276) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 405961989 and 432308570.

Acknowledgements

We would like to thank two anonymous reviewers and the editors for their most helpful comments. Furthermore, we thank Branden Fitelson, Stephan Hartmann, Jan Sprenger, Borut Trpin, Gregory Wheeler and other colleagues at MCMP for their comments and suggestions, as well as audiences at various venues where we presented the paper, the 2019 FEW in Turin, the 2019 BIAP workshop in Girona, and other events in Ancona, Bologna, Cologne, Cork, Edimburgh, Exeter, Groningen, Hannover, London, Sidney, Tilburg, and Vienna.

Barbara Osimani

Polytechnic University of the Marches at Ancona

Department of Biomedical Sciences and Public Health

Via Tronto 10a

Torrette AN

barbaraosimani@gmail.com

Jürgen Landes

Munich Center for Mathematical Philosophy

Ludwigstr. 31

LMU Munich

80539 Munich

Germany

juergen_landesyahoo.de

References

C Glenn Begley and Lee M Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012. URL https://doi.org/10.1038/483531a.

Luc Bovens and Stephan Hartmann. Bayesian Networks and the Problem of Unreliable Instruments. *Philosophy of Science*, 69(1):29–72, 2002. URL https://doi.org/10.1086/338940.

Luc Bovens and Stephan Hartmann. *Bayesian Epistemology*. Oxford University Press, 2003.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013. URL https://doi.org/10.1038/nrn3475.

Lorenzo Casini and Jürgen Landes. Confirmation by Robustness Analysis. A Bayesian Account. forthcoming.

François Claveau and Olivier Grenier. The Variety-of-Evidence Thesis: A Bayesian Exploration of its Surprising Failures. *Synthese*, 196:3001–3028, 2019. URL https://doi.org/10.1007/s11229-017-1607-5.

Franşois Claveau. The Independence Condition in the Variety-of-Evidence Thesis. *Philosophy of Science*, 80(1):94–118, 2013. URL https://doi.org/10.1086/668877.

Open Science Collaboration. Estimating the reproducibility of psychological science. *American Heart Journal*, 349(6251):943–aac4716–8, 2015. URL https://doi.org/10.1126/science.aac4716.

A. Darwiche. *Modeling and reasoning with Bayesian networks*. Cambridge University Press, 2009.

John Earman. *Bayes or Bust?* MIT Press, 1992.

Alexander Etz and Joachim Vandekerckhove. A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE*, 11(2):1–12, 2016. URL https://doi.org/10.1371/journal.pone.0149794.

Branden Fitelson. A bayesian account of independent evidence with applications. *Philosophy of Science*, 68(3):S123–S140, 2001. URL https://doi.org/10.2307/3080940.

Allan Franklin and Colin Howson. Why do scientists prefer to vary their experiments? *Studies in History and Philosophy of Science Part A*, 15(1):51–62, 1984. URL https://doi.org/10.1016/0039-3681(84)90029-3.

Andrew Gelman. Working through some issues. *Significance*, 12(3):33–35, 2015. URL https://doi.org/10.1111/j.1740-9713.2015.00828.x.

Stephan Hartmann and Luc Bovens. The Variety-of-Evidence Thesis and the Reliability of Instruments: A Bayesian-Network Approach, February 2001. URL http://philsci-archive.pitt.edu/235/. last modified: 07 Oct 2010 15:10.

Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, 196:3067–3081, 2019. URL https://doi.org/10.1007/s11229-016-1294-7.

Carl Hempel. *Philosophy of Natural Science*. Prentice Hall, 1966.

Paul Horwich. *Probability and Evidence*. Cambride University Press, 1982.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005. URL https://doi.org/10.1371%2Fjournal.pmed.0020124.

Daniël Lakens and Ellen RK Evers. Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9(3):278–292, 2014.

Jürgen Landes. The Variety of Evidence Thesis and its Independence of Degrees of Independence. forthcoming.

Jürgen Landes. Variety of Evidence. *Erkenntnis*, 85:183–223, 2020. URL https://doi.org/10.1007/s10670-018-0024-6.

Jürgen Landes, Barbara Osimani, and Roland Poellinger. Epistemology of Causal Inference in Pharmacology: Towards a framework for the assessment of harms. *European Journal for Philosophy of Science*, 8:3–49, 2018. URL https://doi.org/10.1007/s13194-017-0169-1.

Maarten Marsman, Felix D Schönbrodt, Richard D Morey, Yuling Yao, Andrew Gelman, and Eric-Jan Wagenmakers. A bayesian bird's eye view of 'replications of important results in social psychology'. *Royal Society Open Science*, 4(1):160426, 2017.

Timothy McGrew. Confirmation, Heuristics, and Explanatory Reasoning. *British Journal for the Philosophy of Science*, 54(4):553–567, 2003. URL https://doi.org/10.1093/bjps/54.4.553.

Paul E. Meehl. Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, 1(2):108–141, 1990. URL https://doi.org/10.1207/s15327965pli0102_1.

Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021, 2017. URL https://doi.org/10.1038/s41562-016-0021.

Richard E. Neapolitan. *Learning Bayesian Networks*. Pearson, 2003.

Erik J. Olsson. *Against Coherence: Truth, Probability, and Justification*. Oxford University Press, 2005.

Barbara Osimani, Mantas Radzvilas, and Francesco De Pretis. Science as a Signaling Game: Statistical Evidence in Strategic Environments. 2020. forthcoming.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 1 edition, 2000.

Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 9(10): 328–329, 2011. URL https://doi.org/10.1038/nrd3439-c1.

Felipe Romero. Can the behavioral sciences self-correct? a social epistemic study. *Studies in History and Philosophy of Science Part A*, 60:55–69, 2016. URL https://doi.org/10.1016/j.shpsa.2016.10.002.

Paul E Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society open science*, 3(9):160384, 2016. URL https://doi.org/10.1098/rsos.160384.

David J Stanley and Jeffrey R Spence. Expectations for Replications: Are Yours Realistic? *Perspectives on Psychological Science*, 9(3):305–318, 2014. URL https://doi.org/10.1177/1745691614528518.

Marcel Weber. *Philosophy of experimental biology*. Cambridge University Press, 2005.

Gregory Wheeler and Richard Scheines. Coherence and Confirmation through Causation. *Mind*, 122(485):135–170, 2013. URL https://doi.org/10.1093/mind/fzt019.

William C. Wimsatt. Robustness, Reliability and Overdetermination. In MB Brewer and BE Colllins, editors, *Scientific Inquiry and the Social Sciences: Festschrift for Donald Campbell*, pages 125–163. Jossey-Bass Publishers, 1981.

William C. Wimsatt. Robustness, Reliability, and Overdetermination (1981). In Léna Soler, Emiliano Trizio, Thomas Nickles, and William Wimsatt, editors, *Characterizing the Robustness of Science*, volume 292 of *Boston Studies in the Philosophy of Science*, pages 61–87. Springer, 2012. URL https://doi.org/10.1007/978-94-007-2759-5_2.

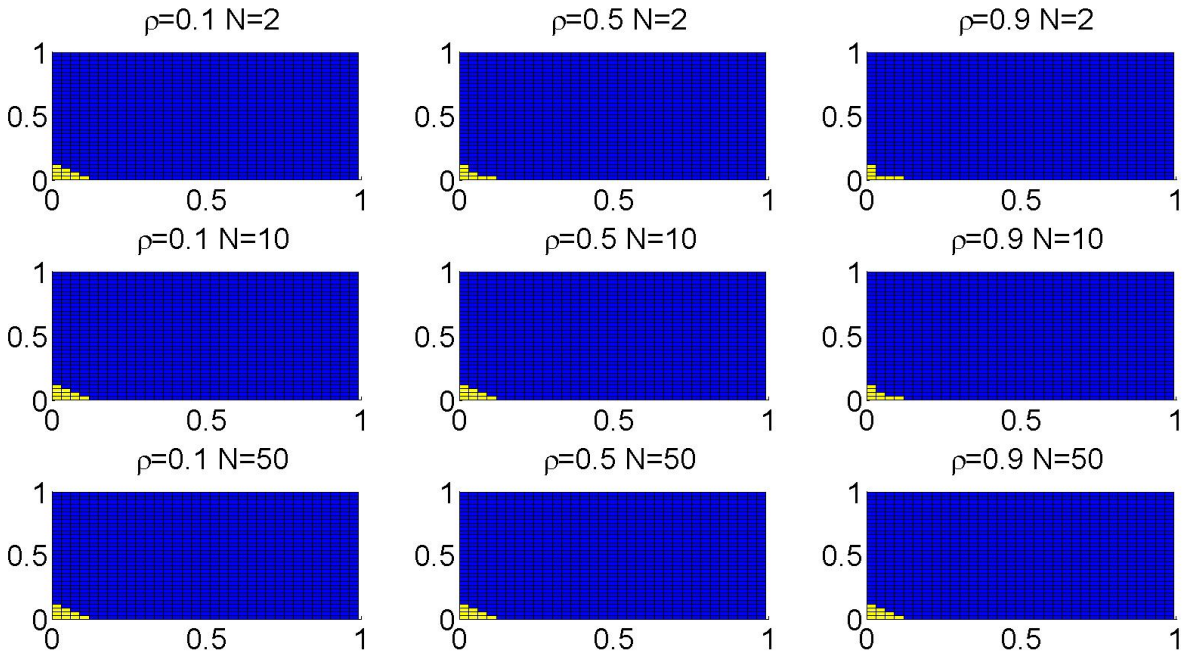

Figure 11: **Scenario 2 for a = 0.1 and varying ρ, N in the p − q-plane**. Yellow indicates the area of VET failure, the blue color means that the VET holds. Figures within one column vary only with the number of reports: top $N = 2$, middle $N = 10$, bottom $N = 50$. Figures within one row vary only with ρ: left $\rho = 0.1$, middle $\rho = 0.5$, right $\rho = 0.9$.

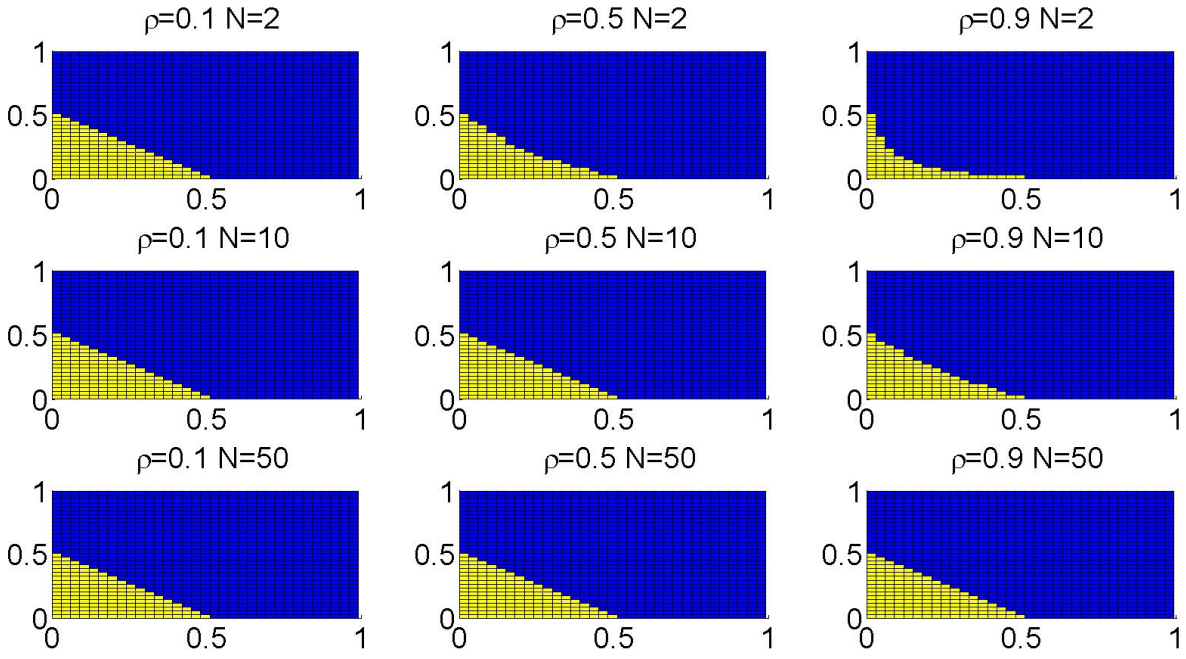

Figure 12: **Scenario 2 for a = 0.5 and varying ρ, N in the p − q-plane**. Yellow indicates the area of VET failure, the blue color means that the VET holds. Figures within one column vary only with the number of reports: top $N = 2$, middle $N = 10$, bottom $N = 50$. Figures within one row vary only with respect to ρ: left $\rho = 0.1$, middle $\rho = 0.5$, right $\rho = 0.9$.

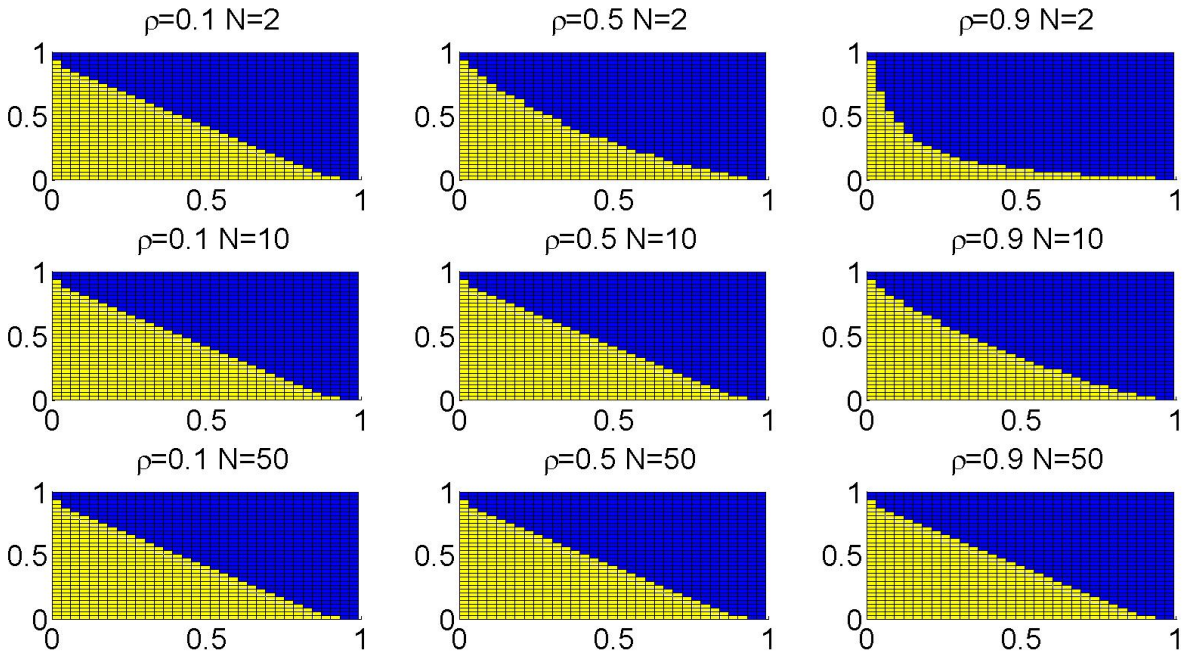

Figure 13: **Scenario 2 for a = 0.9 and varying ρ, N in the p − q-plane**. Yellow indicates the area of VET failure, the blue color means that the VET holds. Figures within one column vary only with the number of reports: top $N = 2$, middle $N = 10$, bottom $N = 50$. Figures within one row vary only with row: left $\rho = 0.1$, middle $\rho = 0.5$, right $\rho = 0.9$.

Appendix A The Variety of Evidence Thesis in the Bovens and Hartmann Framework for Multiple Items of Evidence

A.1 Formal Analysis

The first observation we make is that to determine the difference of two posterior probabilities in the hypothesis of interest being true given the evidence E we only require to compute likelihoods.[28]

Lemma 1. *Under the ceteris paribus assumption of* $P(Hyp) = P_1(Hyp)$

$$\text{sign}(P(Hyp|E) - P_1(Hyp|E)) = \text{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P(E|\overline{Hyp}) \cdot P_1(E|Hyp)\Big) \ .$$

Proof. Applying Bayes' Theorem we find the following equality for all probability functions

[28]Throughout, all probabilities $(P(Hyp), p, 1-p, q, 1-q, a, 1-a, \rho, 1-\rho)$ are non-zero and thus all conditional probabilities are well-defined.

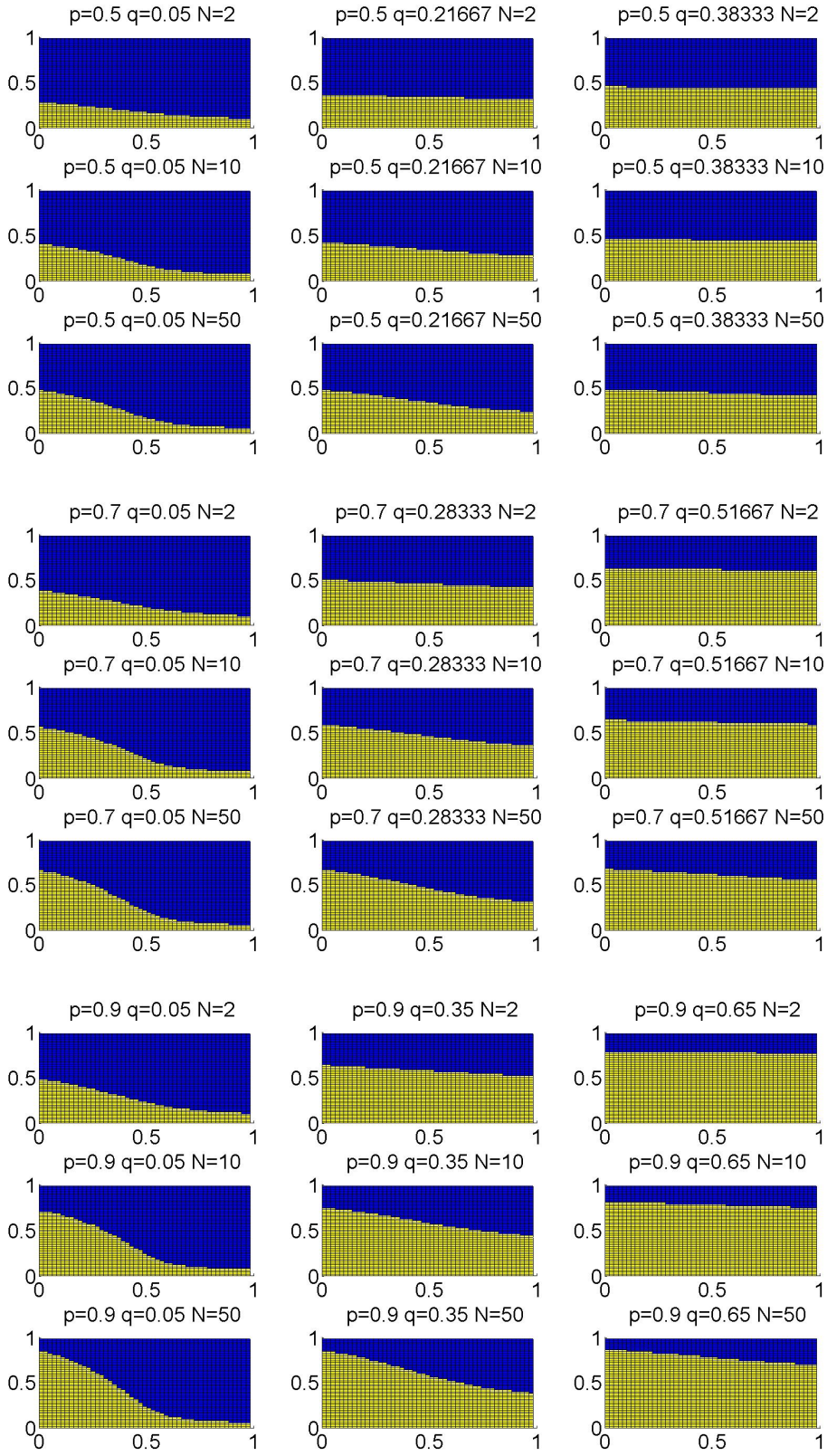

Figure 14: **Scenario 3: The ρ – a-plane for varying p, q, N**. Yellow indicates VET failure. The top $3x3$-set is for $p = 0.5$, the second for $p = 0.7$ and the bottom for $p = 0.9$. Within a column of a $3x3$-set, N varies: top $N = 2$, middle $N = 10$, bottom $N = 50$. Within a row q varies: left $q = 0.05$, middle $q = 0.05 + p/3$, right $q = 0.05 + 2p/3$.

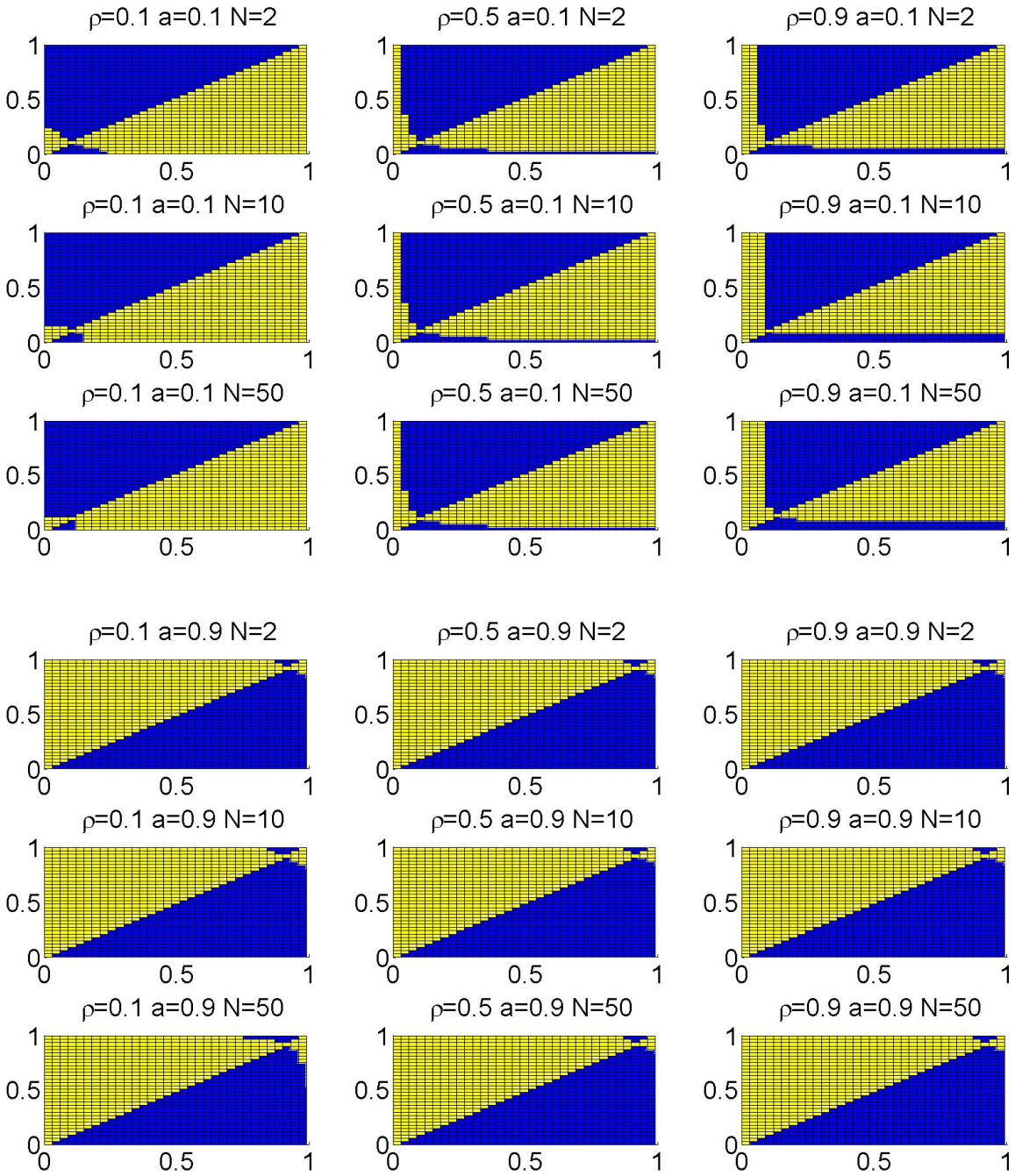

Figure 15: **Scenario 3: The p, q-plane for varying ρ, a, N**. Yellow indicates the area of VET failure, blue means that the VET holds. The first $3x3$-set is for $a = 0.1$ and the bottom for $a = 0.9$. Within a column of a $3x3$-set, N varies: top $N = 2$, middle $N = 10$, bottom $N = 50$. Within a row ρ varies: left $\rho = 0.1$, middle $\rho = 0.5$, right $\rho = 0.9$.

P for which the conditional probabilities are well-defined

$$P(Hyp|E) = \frac{P(Hyp)}{P(Hyp) + P(\overline{Hyp})\frac{P(E|\overline{Hyp})}{P(E|Hyp)}} \quad .$$

Hence,

$$\begin{aligned}\text{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big) &= \text{sign}\Big(\frac{P_1(E|\overline{Hyp})}{P_1(E|Hyp)} - \frac{P(E|\overline{Hyp})}{P(E|Hyp)}\Big) \\ &= \text{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P(E|\overline{Hyp}) \cdot P_1(E|Hyp)\Big) \ .\end{aligned}$$

□

So, whenever we want to compute which of two posterior probabilities is greater, $P(Hyp|E), P_1(Hyp|E)$, we only need to compute the terms in Lemma 1.

Some results in this subsection also appear in the Philsci-Archive in Hartmann and Bovens (2001). Bovens and Hartmann never published the results in their Philsci-report elsewhere. Our and their results are consistent and were found independently.

A.1.1 Scenario 1

We first compare two situations in which we have a number of reports pertaining to a consequence of the hypothesis. In the first situation, the reliability of the reports are independent, in the second situation reliability is shared, see first scenario in Figure 2 for the Bayesian network representations. We use E^+ to denote the number of positive reports and E^- for the negative reports.

The ceteris paribus conditions we impose are those in Bovens & Hartmann. P and P_1 agree on their joint domain: $P(Hyp) = P_1(Hyp)$, $P(Con|Hyp) = P_1(Con|Hyp) =: p$ and $P(\overline{Con}|Hyp) = P_1(\overline{Con}|Hyp) =: q$ as well as

$$\begin{aligned}P(Rep_i|Con, Rel_i) &= P_1(Rep|Con, Rel) = 1 \text{ for all } i \\ P(Rep_i|\overline{Con}, Rel_i) &= P_1(Rep|\overline{Con}, Rel) = 0 \text{ for all } i \\ P(Rep_i|Con, \overline{Rel_i}) &= P_1(Rep|Con, \overline{Rel}) = a \text{ for all } i \\ P(Rep_i|\overline{Con}, \overline{Rel_i}) &= P_1(Rep|\overline{Con}, \overline{Rel}) = a \text{ for all } i \\ P(Rel_i) &= P_1(Rel) = \rho \text{ for all } i \ .\end{aligned}$$

Proposition 1. *In Scenario 1 it holds for all $E^+ \geq 1$-many positive reports that*

$$\text{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big) = \text{sign}\Big((\rho + \bar{\rho} \cdot a)^{E^+} - (\bar{\rho}^{E^+-1} \cdot \rho + (\bar{\rho} \cdot a)^{E^+})\Big) \ .$$

Proof. Let us compute conditional probabilities

$$\begin{aligned}
P(E|Hyp) =& p \cdot \Big(\sum_{g=0}^{E^+} \binom{E^+}{g} \cdot \rho^g \cdot \bar{\rho}^{E^+-g} \cdot a^{E^+-g}\Big) + \bar{p} \cdot \Big(\bar{\rho}^{E^+} \cdot a^{E^+}\Big) \\
=& p \cdot \Big(\rho + \bar{\rho} \cdot a\Big)^{E^+} + \bar{p} \cdot \Big(\bar{\rho}^{E^+} \cdot a^{E^+}\Big) \\
P(E|\overline{Hyp}) =& q \cdot \Big(\sum_{g=0}^{E^+} \binom{E^+}{g} \cdot \rho^g \cdot \bar{\rho}^{E^+-g} \cdot a^{E^+-g}\Big) + \bar{q} \cdot \Big(\bar{\rho}^{E^+} \cdot a^{E^+}\Big) \\
=& q \cdot \Big(\rho + \bar{\rho} \cdot a\Big)^{E^+} + \bar{q} \cdot \Big(\bar{\rho}^{E^+} \cdot a^{E^+}\Big) \\
P_1(E|Hyp) =& \bar{\rho} \cdot a^{E^+} + \rho \cdot p \\
P_1(E|\overline{Hyp}) =& \bar{\rho} \cdot a^{E^+} + \rho \cdot q \\
& \text{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P_1(E|Hyp) \cdot P(E|\overline{Hyp})\Big) \\
& = \text{sign}\Big((\bar{\rho} \cdot a^{E^+} + \rho \cdot q) \cdot \Big(p \cdot (\rho + \bar{\rho} \cdot a)^{E^+} + \bar{p} \cdot (\bar{\rho}^{E^+} \cdot a^{E^+})\Big) \\
& \quad - (\bar{\rho} \cdot a^{E^+} + \rho \cdot p) \cdot \Big(q \cdot (\rho + \bar{\rho} \cdot a)^{E^+} + \bar{q} \cdot (\bar{\rho}^{E^+} \cdot a^{E^+})\Big)\Big) \\
& = \text{sign}\Big(\bar{\rho} \cdot a^{E^+} \cdot (\rho + \bar{\rho} \cdot a)^{E^+}(p - q) + (\bar{\rho} \cdot a)^{E^+} \cdot [\bar{p} \cdot (q \cdot \rho + a^{E^+}\bar{\rho}) - \bar{q} \cdot (p \cdot \rho + a^{E^+}\bar{\rho})]\Big) \\
& = \text{sign}\Big(\bar{\rho} \cdot a^{E^+} \cdot (\rho + \bar{\rho} \cdot a)^{E^+}(p - q) + (\bar{\rho} \cdot a)^{E^+} \cdot [(q - p) \cdot \rho + a^{E^+} \cdot \bar{\rho} \cdot (q - p)]\Big) \\
& = \text{sign}\Big(\bar{\rho} \cdot a^{E^+} \cdot (p - q) \cdot [(\rho + \bar{\rho} \cdot a)^{E^+} - (\bar{\rho}^{E^+-1} \cdot \rho + (\bar{\rho} \cdot a)^{E^+})]\Big) \ .
\end{aligned}$$

Since $p > q$, the leading factors are strictly greater than zero and do not influence the sign of the expression. □

It follows immediately, that if $\rho = \bar{\rho} = 0.5$ and $a = 0$, then both posterior probabilities are equal. More generally, we have

Proposition 2. *The status of the VET is independent of p, q but does depend on the relation of ρ, a, E^+:*

- *if $\frac{\rho}{\bar{\rho}} + a < 1$, then the VET fails for all $E^+ \geq \frac{\log(\frac{\rho}{\bar{\rho}})}{\log(a + \frac{\rho}{\bar{\rho}})} > 1$.*

- *If* $\frac{\rho}{\bar{\rho}} + a \geq 1$*, then the VET holds for all* $E^+ \geq 2$*.*

Proof. Let $\lambda := \frac{\rho}{\bar{\rho}}$, in particular, assume that it is well-defined, i.e., $\bar{\rho} > 0$. With this new parameter we find

$$\begin{aligned}(\rho + \bar{\rho} \cdot a)^{E^+} - (\bar{\rho}^{E^+-1} \cdot \rho + (\bar{\rho} \cdot a)^{E^+}) =& (\lambda + a)^{E^+} \cdot \bar{\rho}^{E^+} - \lambda \cdot \bar{\rho}^{E^+} - a^{E^+} \cdot \bar{\rho}^{E^+} \\ =& \bar{\rho}^{E^+} \cdot \left((\lambda + a)^{E^+} - (\lambda + a^{E^+})\right) .\end{aligned}$$

If $\lambda + a \geq 1$, then we find for $E^+ \geq 2$ that the VET holds:

$$\begin{aligned}(\rho + \bar{\rho} \cdot a)^{E^+} - (\bar{\rho}^{E^+-1} \cdot \rho + (\bar{\rho} \cdot a)^{E^+}) =& \bar{\rho}^{E^+} \cdot \left((\lambda + a)^{E^+} - (\lambda + a^{E^+})\right) \\ >& \bar{\rho}^{E^+} \cdot \left((\lambda + a)^{E^+} - (\lambda + a)\right) \\ >& \bar{\rho}^{E^+} \cdot \left((\lambda + a) - \lambda - a\right) \\ =& 0 \ .\end{aligned}$$

For $\lambda + a < 1$ we find

$$\begin{aligned}(\rho + \bar{\rho} \cdot a)^{E^+} - (\bar{\rho}^{E^+-1} \cdot \rho + (\bar{\rho} \cdot a)^{E^+}) =& \bar{\rho}^{E^+} \cdot \left((\lambda + a)^{E^+} - (\lambda + a^{E^+})\right) && \text{(A.1)} \\ <& \bar{\rho}^{E^+} \cdot \left((\lambda + a)^{E^+} - \lambda\right) \ . && \text{(A.2)}\end{aligned}$$

Since $(\lambda+a)^{E^+}$ tends to zero as E^+ gets ever greater, the VET fails for large enough E^+. Solving $(\lambda + a)^{E^+} - \lambda = 0$ for E^+ we find the threshold of $E^+ = \frac{\log(\lambda)}{\log(\lambda+a)}$ beyond which the VET fails. □

Let us next note that the condition $\frac{\rho}{\bar{\rho}} + a < 1$ is equivalent to $a < \frac{1-2\rho}{1-\rho}$. Since the inequality in (A.2) in the proof of this proposition is not tight, the VET may already fail for smaller e than this bound. Since solving (A.1)$= 0$ for E^+ can only be done numerically, the curves $(\lambda + a)^{E^+} - \lambda = (\frac{\rho}{\bar{\rho}} + a)^{E^+} - \frac{\rho}{\bar{\rho}} = 0$ are plotted for various $E^+ > 2$. We can see the area in which the VET fails in the $\rho - a$-plane grow as the number of positive reports increases, see Figure 3.

A.1.2 Scenario 2

We now turn to the second scenario in which a single positive report is, in certain cases, more confirmatory than a number of positive reports, ceteris paribus.

The ceteris paribus conditions we impose are the usual ones: $P(Hyp) = P_1(Hyp), P(Con_n|Hyp) = P_1(Con|Hyp) = p$ for all n, $P(\overline{Con_n}|Hyp) = P_1(\overline{Con}|Hyp) = q$ for all n, $P(Rel) = P_1(Rel) = \rho$ and

$$
\begin{aligned}
&P(Rep_i|Con_n, Rel) = P_1(Rep|Con, Rel) = 1 \text{ for all } i, n \\
&P(Rep_i|\overline{Con_n}, Rel) = P_1(Rep|\overline{Con_n}, Rel) = 0 \text{ for all } i, n \\
&P(Rep_i|Con_n, \overline{Rel}) = P_1(Rep|Con, \overline{Rel}) = a \text{ for all } i, n \\
&P(Rep_i|\overline{Con_n}, \overline{Rel}) = P_1(Rep|\overline{Con}, \overline{Rel}) = a \text{ for all } i, n \ .
\end{aligned}
$$

Let us first note that discordant evidence entails that the source of evidence in Scenario 1 is unreliable. Hence, $P_1(Hyp|Rep) > P_1(Hyp) = P(Hyp)$. For a discordant body of evidence E we have that $P_1(Hyp|E) > P_1(Hyp) = P(Hyp)$. Let us hence assume that all reports are positive. To determine the status of the VET we find for N positive reports:

Proposition 3. *For all $N \geq 2$ it holds in Scenario 2 that*

$$
\begin{aligned}
&\text{sign}\Big(P(Hyp|Rep_1, \ldots, Rep_N) - P_1(Hyp|Rep)\Big) \\
&\quad = \text{sign}\Big((\rho \cdot q \cdot p + \bar{\rho} \cdot a(p+q)) \cdot (p+q)^{N-2} - \bar{\rho} \cdot a^N\Big) \ .
\end{aligned}
$$

Proof. We find

$$
\begin{aligned}
P(Rep_1, \ldots, Rep_N|Hyp) &= \sum_{CON_1, \ldots, CON_N, REL} P(Rep_1, \ldots, Rep_N, CON_1, \ldots, CON_N, REL|Hyp) \\
&= \rho \cdot p^N + \bar{\rho} \cdot (a \cdot \rho + a \cdot \bar{\rho})^N = \rho \cdot p^N + \bar{\rho} \cdot a^N \\
P(Rep_1, \ldots, Rep_N|\overline{Hyp}) &= \sum_{CON_1, \ldots, CON_N, REL} P(Rep_1, \ldots, Rep_N, CON_1, \ldots, CON_N, REL|\overline{Hyp}) \\
&= \rho \cdot q^N + \bar{\rho} \cdot (a \cdot \rho + a \cdot \bar{\rho})^N = \rho \cdot q^N + \bar{\rho} \cdot a^N
\end{aligned}
$$

$$
\begin{aligned}
P_1(Rep|Hyp) &= \sum_{CON,REL} P_1(Rep, CON, REL|Hyp) \\
&= p \cdot (\rho + a \cdot \bar{\rho}) + \bar{p} a \bar{\rho} = p \cdot \rho + a \cdot \bar{\rho} \\
P_1(Rep|\overline{Hyp}) &= \sum_{CON,REL} P_1(Rep, CON, REL|\overline{Hyp}) \\
&= q \cdot (\rho + a \cdot \bar{\rho}) + \bar{q} a \bar{\rho} = q \cdot \rho + a \cdot \bar{\rho} \ .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\text{sign}\Big(P(Hyp|Rep_1, \ldots, Rep_N) - P_1(Hyp|Rep)\Big) \\
&= \text{sign}\Big(\frac{P_1(Rep^+|\overline{Hyp})}{P_1(Rep^+|Hyp)} - \frac{P(Rep_1^+, \ldots, Rep_N^+|\overline{Hyp})}{P(Rep_1^+, \ldots, Rep_N^+)|Hyp}\Big) \\
&= \text{sign}\Big(\frac{q \cdot \rho + a \cdot \bar{\rho}}{p \cdot \rho + a \cdot \bar{\rho}} - \frac{\rho \cdot q^N + a^N \cdot \bar{\rho}}{\rho \cdot p^N + a^N \cdot \bar{\rho}}\Big) \\
&= \text{sign}\Big(\rho^2 \cdot q \cdot p \cdot (p^{N-1} - q^{N-1}) + a^N \cdot \rho \cdot \bar{\rho} \cdot (q - p) + a \cdot \bar{\rho} \cdot \rho \cdot (p^N - q^N)\Big) \\
&= \text{sign}\Big(\rho^2 \cdot q \cdot p \cdot (p^{N-1} - q^{N-1}) + \rho \cdot \bar{\rho} \cdot a \cdot [-a^{N-1} \cdot (p - q) + (p^N - q^N)]\Big) \\
&= \text{sign}\Big(\rho^2 \cdot q \cdot p \cdot (p^{N-1} - q^{N-1}) + \rho \cdot \bar{\rho} \cdot a \cdot (p - q) \cdot [-a^{N-1} + (p + q)^{N-1}]\Big) \\
&= \text{sign}\Big(\rho \cdot (p - q) \cdot [\rho \cdot q \cdot p \cdot (p + q)^{N-2} + \bar{\rho} \cdot a[-a^{N-1} + (p + q)^{N-1}]\Big) \ .
\end{aligned}
$$

□

Note that $p + q > a$ is a sufficient condition for this sign being positive and the VET holding.

For ever larger N the condition $p + q > a$ eventually becomes a necessary condition, too. This can be seen from the last expression by substituting $\lambda \cdot a$ for $(p + q)$ with $\lambda \cdot a \in (p + q, a)$. We obtain

$$
(\rho \cdot q \cdot p + \bar{\rho} \cdot a^2 \lambda) \cdot (a\lambda)^{N-2} - \bar{\rho} \cdot a^N = a^{N-2} \cdot [(\rho \cdot q \cdot p + \bar{\rho} \cdot a^2 \lambda) \cdot \lambda^{N-2} - \bar{\rho} \cdot a^2]
$$

Since only if $\lambda < 1$, is the term on the left of the minus sign eventually less than the term on the right of it.

This means that the area where the VET fails grows when N is large. It should be noted that the values of ρ has next to no influence on whether the VET fails or holds for large N. See

Figure 21 and Figure 22 for graphical illustrations.

A.1.3 Scenario 3

We now investigate the third scenario. Again, it is the single source giving multiple reports which provides, in some cases, more confirmation to the hypothesis than multiple sources, ceteris paribus.

The ceteris paribus conditions we here impose are: $P(Hyp) = P_1(Hyp)$, $P(Con_n|Hyp) = P_1(Con_n|Hyp) = p$ for all n, $P(\overline{Con_n}|Hyp) = P_1(\overline{Con_n}|Hyp) = q$ for all n, $P(Rel) = P_1(Rel_i) = \rho$ for all i and

$$
\begin{aligned}
P(Rep_i|Con_i, Rel) &= P_1(Rep_i|Con_i, Rel_i) = 1 \text{ for all } i \\
P(Rep_i|\overline{Con_i}, Rel) &= P_1(Rep|\overline{Con_i}, Rel_i) = 0 \text{ for all } i \\
P(Rep_i|Con_i, \overline{Rel}) &= P_1(Rep|Con_i, \overline{Rel_i}) = a \text{ for all } i \\
P(Rep_i|\overline{Con_i}, \overline{Rel}) &= P_1(Rep|\overline{Con_i}, \overline{Rel_i}) = a \text{ for all } i \ .
\end{aligned}
$$

Even when all reports are positive, the general formula for the difference of conditional probabilities is rather opaque:

Proposition 4. *For all $E^+ \geq 1$ it holds in Scenario 3 that*

$$
\begin{aligned}
&\text{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big) \\
&= \text{sign}\Big(\bar{\rho} \cdot a^{E^+} \cdot [\sum_{f=1}^{E^+} \binom{E^+}{f} \rho^f \cdot (\bar{\rho} \cdot a)^{E^+ - f} \cdot (q^f - p^f)] \\
&\quad + \rho \cdot [\sum_{k=0}^{E^+ - 1} \binom{E^+}{k} \cdot \rho^k \cdot (\bar{\rho} \cdot a)^{E^+ - k} \cdot (q^k \cdot p^{E^+} - p^k \cdot q^{E^+})]\Big) \ .
\end{aligned}
$$

Proof. For P and P_1 we calculate and find

$$
\begin{aligned}
P(E|Hyp) &= \sum_{CON_1, \ldots, CON_{E^+}, REL} P(E, CON_1, \ldots, CON_{E^+}, REL|Hyp) \\
&= \rho \cdot p^{E^+} + \bar{\rho} \cdot (a \cdot \rho + a \cdot \bar{\rho})^{E^+} = \rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}
\end{aligned}
$$

$$
\begin{aligned}
P(E|\overline{Hyp}) =& \sum_{CON_1,\ldots,CON_{E^+},REL} P(E, CON_1, \ldots, CON_{E^+}, REL|\overline{Hyp}) \\
=& \rho \cdot q^{E^+} + \bar{\rho} \cdot (a \cdot \rho + a \cdot \bar{\rho})^{E^+} = \rho \cdot q^{E^+} + \bar{\rho} \cdot a^{E^+} \\
P_1(E|\overline{Hyp}) =& \sum_{CON_1,\ldots,CON_{E^+},REL_1,\ldots,REL_{E^+}} P_1(E, CON_1, \ldots, CON_{E^+}, REL_1, \ldots, REL_{E^+}|\overline{Hyp}) \\
=& (q \cdot (\rho + \bar{\rho} \cdot a) + (1-q) \cdot \bar{\rho} \cdot a)^{E^+} = (q \cdot \rho + \bar{\rho} \cdot a)^{E^+} \\
P_1(E|Hyp) =& \sum_{CON_1,\ldots,CON_{E^+},REL_1,\ldots,REL_{E^+}} P_1(E, CON_1, \ldots, CON_{E^+}, REL_1, \ldots, REL_{E^+}|Hyp) \\
=& (p \cdot (\rho + \bar{\rho} \cdot a) + (1-p) \cdot \bar{\rho} \cdot a)^{E^+} = (p \cdot \rho + \bar{\rho} \cdot a)^{E^+} \quad .
\end{aligned}
$$

We find the following expression which is symmetric in p and q.

$$
\begin{aligned}
&\operatorname{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P_1(E|Hyp) \cdot P(E|\overline{Hyp})\Big) \\
&= \operatorname{sign}\Big((q \cdot \rho + \bar{\rho} \cdot a)^{E^+} \cdot (\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}) - (p \cdot \rho + \bar{\rho} \cdot a)^{E^+} \cdot (\rho \cdot q^{E^+} + \bar{\rho} \cdot a^{E^+})\Big) \\
&= \operatorname{sign}\Big(\bar{\rho} \cdot a^{E^+} \cdot [(q \cdot \rho + \bar{\rho} \cdot a)^{E^+} - (p \cdot \rho + \bar{\rho} \cdot a)^{E^+}] \\
&\quad + (q \cdot \rho + \bar{\rho} \cdot a)^{E^+} \cdot (\rho \cdot p^{E^+}) - (p \cdot \rho + \bar{\rho} \cdot a)^{E^+} \cdot (\rho \cdot q^{E^+})\Big) \\
&= \operatorname{sign}\Big(\bar{\rho} \cdot a^{E^+} \cdot [\sum_{f=1}^{E^+} \binom{E^+}{f} \rho^f \cdot (\bar{\rho} \cdot a)^{E^+-f} \cdot (q^f - p^f)] \\
&\quad + \rho \cdot [\sum_{k=0}^{E^+-1} \binom{E^+}{k} \cdot \rho^k \cdot (\bar{\rho} \cdot a)^{E^+-k} \cdot (q^k \cdot p^{E^+} - p^k \cdot q^{E^+})]\Big) \quad .
\end{aligned}
$$

□

To illuminate the situation, we consider special cases. At first we consider the $p-q-a$-space for fixed $\rho = 0.5$:

Proposition 5. *For $\rho = \frac{1}{2}$ and $E^+ \geq 2$ the VET fails, if and only if*

$$
p \cdot q \geq a^2 \quad .
$$

Proof.

$$
\operatorname{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P_1(E|Hyp) \cdot P(E|\overline{Hyp})\Big)
$$

$$= \text{sign}\Big((q+a)^{E^+} \cdot (p^{E^+} + a^{E^+}) - (p+a)^{E^+} \cdot (q^{E^+} + a^{E^+})\Big) \ .$$

Letting $\lambda := \frac{p}{a}$ and $\mu := \frac{q}{a}$ this simplifies to

$$\begin{aligned}
&\text{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P_1(E|Hyp) \cdot P(E|\overline{Hyp})\Big) \\
&= \text{sign}\Big(\Big(\frac{q+a}{p+a}\Big)^{E^+} - \frac{q^{E^+} + a^{E^+}}{p^{E^+} + a^{E^+}}\Big) \\
&= \text{sign}\Big(\Big(\frac{a \cdot \mu + a}{a \cdot \lambda + a}\Big)^{E^+} - \frac{(a \cdot \mu)^{E^+} + a^{E^+}}{(a \cdot \lambda)^{E^+} + a^{E^+}}\Big) \\
&= \text{sign}\Big(\Big(\frac{1+\mu}{1+\lambda}\Big)^{E^+} - \frac{\mu^{E^+} + 1}{\lambda^{E^+} + 1}\Big) \\
&= \text{sign}\Big((1+\mu)^{E^+} \cdot (\lambda^{E^+} + 1) - (1+\lambda)^{E^+} \cdot (\mu^{E^+} + 1)\Big) \\
&= \text{sign}\Big((\lambda^{E^+} + 1) \cdot [\sum_{i=0}^{E^+} \binom{E^+}{i} \mu^i] - (\mu^{E^+} + 1) \cdot [\sum_{i=0}^{E^+} \binom{E^+}{i} \lambda^i]\Big) \\
&= \text{sign}\Big(\cdot \sum_{i=0}^{E^+} \binom{E^+}{i} [(\lambda^{E^+} + 1)\mu^i - (\mu^{E^+} + 1) \cdot \lambda^i]\Big) \\
&= \text{sign}\Big(\sum_{i=0}^{E^+} \binom{E^+}{i} \cdot [(\mu^i - \lambda^i) + (\mu \cdot \lambda)^i \cdot (\lambda^{E^+-i} - \mu^{E^+-i})]\Big) \\
&= \text{sign}\Big(\sum_{i=0}^{E^+} \binom{E^+}{i} \cdot [(\mu^i - \lambda^i) + (\mu \cdot \lambda)^{E^+-i} \cdot (\lambda^i - \mu^i)]\Big) \\
&= \text{sign}\Big(\sum_{i=0}^{E^+} \binom{E^+}{i} \cdot (\mu^i - \lambda^i) \cdot (1 - (\mu \cdot \lambda)^{E^+-i})\Big) \ .
\end{aligned}$$

For the second to last step we used that $\binom{E^+}{i} = \binom{E^+}{E^+-i}$.

Since $\mu^i - \lambda^i < 0$ for all $i > 0$ and equal to zero for $i = 0$ and since $E^+ \geq 2$, this expression is zero, if and only if $1 = \mu \cdot \lambda$ which holds, if and only if $p \cdot q = a^2$. If $p \cdot q > a^2$, then $1 - (\mu \cdot \lambda)^i < 0$ for $i \geq 1$ and hence the polynomial expression is positive, as required. □

This generalises the result of (Bovens and Hartmann 2003, p. 102-103) simultaneously to arbitrary $a \in (0, 1)$ and to arbitrary $E^+ \geq 2$. If $p \cdot q < a^2$, then the VET holds and for $p \cdot q = a^2$ the two posteriors are equal. Note that since $p > q$, the VET fails for $q > a$, since then $p \cdot q > a^2$.

Finally, we turn to the $a - \rho$-plane. Let us thus fix $0 < q < p < 1$ and vary a and ρ. We easily find

Proposition 6. *For all* $0 < q < p < a < 1$ *and all* $0 < \rho < 1$ *there exists there exists some* $N \in \mathbb{N}$ *such that for all* $E^+ \geq N$ *the VET holds.*

Proof.

$$\begin{aligned}
&\operatorname{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P_1(E|Hyp) \cdot P(E|\overline{Hyp})\Big) \\
&= \operatorname{sign}\Big((q \cdot \rho + \bar{\rho} \cdot a)^{E^+} \cdot (\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}) - (p \cdot \rho + \bar{\rho} \cdot a)^{E^+} \cdot (\rho \cdot q^{E^+} + \bar{\rho} \cdot a^{E^+})\Big) \\
&= \operatorname{sign}\Big(\Big(\frac{q \cdot \rho + \bar{\rho} \cdot a}{p \cdot \rho + \bar{\rho} \cdot a}\Big)^{E^+} - \frac{\rho \cdot q^{E^+} + \bar{\rho} \cdot a^{E^+}}{\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}}\Big) \ .
\end{aligned}$$

The first term of this expression, $\big(\frac{q \cdot \rho + \bar{\rho} \cdot a}{p \cdot \rho + \bar{\rho} \cdot a}\big)^{E^+}$, is decreasing in E^+ with limit zero, since $p > q$. The fate of the VET – in the limit – can be determined by the second term, $-\frac{\rho \cdot q^{E^+} + \bar{\rho} \cdot a^{E^+}}{\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}}$.

Let us next observe that

$$-\frac{\rho \cdot q^{E^+} + \bar{\rho} \cdot a^{E^+}}{\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}} = -\frac{\rho \cdot q^{E^+}}{\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}} - \frac{\bar{\rho} \cdot a^{E^+}}{\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}} \ .$$

Clearly, the first summand on the right of the equality symbol converges to zero, $q < p$.

For the second summand we find

$$-\frac{\bar{\rho} \cdot a^{E^+}}{\rho \cdot p^{E^+} + \bar{\rho} \cdot a^{E^+}} = -\frac{\bar{\rho}}{\rho \cdot \frac{p^{E^+}}{a^{E^+}} + \bar{\rho}} \ .$$

If $a > p$, then the denominator converges to $\bar{\rho}$. Hence, the entire term converges to -1. This means that for all $0 < q < p < a < 1$ and all $0 < \rho < 1$ there exists some $N \in \mathbb{N}$ such that for all $E^+ \geq N$ the VET holds. □

Note that the value of $\rho \in (0, 1)$ plays no role in this statement. We can see this in Figure 4 where for $a > p = 0.9$ the VET holds. Proposition 6 generalises the region where the VET holds to arbitrary N and almost all ρ.

It is tempting to conjecture that for $a < q$, the VET fails, no matter the value of ρ and the value of $q < p < 1$. Unfortunately, this has to be left to further analysis.

Appendix B Appendix for the Proposed Model

B.1 Formal Analysis

B.1.1 Scenario 1

Proposition 7. *For $N = 2$ positive reports we have*

$$\begin{aligned}&\operatorname{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big)\\ &= \operatorname{sign}\Big((\rho(1-\epsilon_+) + (1-\rho)\alpha)^2 \cdot (\rho\epsilon_-^2 + (1-\rho)\gamma^2) - (\rho(1-\epsilon_+)^2 + (1-\rho)\alpha^2) \cdot (\rho\epsilon_- + (1-\rho)\gamma)^2\Big)\ .\end{aligned}$$

Proof. As usual, we only need to calculate likelihoods to determine the fate of the VET. We begin by calculating them separately, using $T_1, \ldots, T_4$ as abbreviations and $b = 1 - \rho$

$$\begin{aligned}P(E|Hyp) =& \sum_{REL_1,\ldots,REL_N} p \cdot P(E, REL_1, \ldots, REL_N|Hyp, Con)\\ &+ \bar{p} \cdot P(E, REL_1, \ldots, REL_N|Hyp, \overline{Con}) =: p \cdot T_1 + \bar{p} \cdot T_2\\ =& \sum_{REL_1,\ldots,REL_N} p \prod_{n=1}^{N} P(Rep_n|Con, REL_n) \cdot P(REL_n)\\ &+ \sum_{REL_1,\ldots,REL_N} \bar{p} \prod_{n=1}^{N} P(Rep_n|\overline{Con}, REL_n) \cdot P(REL_n)\\ =& p \cdot \Big((\rho(1-\epsilon_+))^2 + (b\alpha)^2 + \rho(1-\epsilon_+) \cdot [b\alpha] + b\alpha \cdot [\rho(1-\epsilon_+)]\Big)\\ &+ \bar{p} \cdot \Big((\rho\epsilon_-)^2 + (b\gamma)^2 + \rho\epsilon_- \cdot [b\gamma] + b\gamma \cdot [\rho\epsilon_-]\Big)\ .\end{aligned}$$

$$\begin{aligned}P(E|\overline{Hyp}) =& \sum_{REL_1,\ldots,REL_N} q \cdot P(E, REL_1, \ldots, REL_N|Hyp, Con)\\ &+ \bar{q} \cdot P(E, REL_1, \ldots, REL_N|Hyp, \overline{Con}) = qT_1 + \bar{q}T_2\ .\end{aligned}$$

Let us now investigate P_1:

$$P_1(E|Hyp) = p \sum_{REL} \cdot P_1(REL) \prod_{n=1}^{2} P_1(Rep_i|Con, REL)$$

$$
\begin{aligned}
&\quad + \bar{p} \sum_{REL} \cdot P_1(REL) \prod_{n=1}^{2} P_1(Rep_i|\overline{Con}, REL) \\
&=: pT_3 + \bar{p}T_4 \\
&= p \cdot (\rho(1-\epsilon_+)^2 + b\alpha^2) + \bar{p} \cdot (\rho\epsilon_-^2 + b\gamma^2) \\
P_1(E|\overline{Hyp}) &= qT_3 + \bar{q}T_4 \ .
\end{aligned}
$$

To determine whether the VET holds, we use Lemma 1 and $p > q$ (hence $\bar{q} > \bar{p}$) to find

$$
\begin{aligned}
&\text{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big) \\
&\quad = \text{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P(E|\overline{Hyp}) \cdot P_1(E|Hyp)\Big) \\
&\quad = \text{sign}\Big((pT_1 + \bar{p}T_2) \cdot (qT_3 + \bar{q}T_4) - (qT_1 + \bar{q}T_2) \cdot (pT_3 + \bar{p}T_4)\Big) \\
&\quad = \text{sign}\Big(p\bar{q}(T_1T_4 - T_2T_3) - \bar{p}q(T_2T_3 - T_1T_4)\Big) \\
&\quad = \text{sign}\Big((p\bar{q} - \bar{p}q)(T_1T_4 - T_2T_3)\Big) \qquad \text{(B.1)} \\
&\quad = \text{sign}\Big(T_1T_4 - T_2T_3\Big) \qquad \text{(B.2)} \\
&\quad = \text{sign}\Big((\rho(1-\epsilon_+) + (1-\rho)\alpha)^2 \cdot (\rho\epsilon_-^2 + (1-\rho)\gamma^2) \\
&\qquad\quad - (\rho(1-\epsilon_+)^2 + (1-\rho)\alpha^2) \cdot (\rho\epsilon_- + (1-\rho)\gamma)^2\Big) \ .
\end{aligned}
$$

□

Proposition 8. *$P(Hyp|E) - P_1(Hyp|E) = 0$, if and only if one of the following conditions holds*

$$
\gamma = \frac{\epsilon_-}{1-\epsilon_+} \cdot \alpha \qquad\qquad \gamma = \frac{\epsilon_- \cdot [2\rho(1-\epsilon_+) + \alpha(1-2\rho)]}{(2\rho-1)(1-\epsilon_+) + 2\alpha(1-\rho)} \ .
$$

Proof. We first check that everything is well-defined, i.e., $(2\rho-1)(1-\epsilon_+) + 2\alpha(1-\rho) \neq 0$. Noting that $\alpha > 1 - \epsilon_+$ we find

$$
\begin{aligned}
(2\rho-1)(1-\epsilon_+) + 2\alpha(1-\rho) &> (2\rho-1)(1-\epsilon_+) + 2(1-\epsilon_+)(1-\rho) \\
&= (1-\epsilon_+) \cdot [2\rho - 1 + 2 - 2\rho] = (1-\epsilon_+) \\
&> 0 \ .
\end{aligned}
$$

The expression $T_1T_4 - T_2T_3$ can be understood as polynomial in the variable γ of degree two with parameters, $\alpha, \rho, \epsilon_+, \epsilon_-$. We proceed to find the roots of this polynomial. We guess that

$$\begin{aligned}&T_1T_4 - T_2T_3\\ =&(\rho(1-\epsilon_+) + (1-\rho)\alpha)^2 \cdot (\rho\epsilon_-^2 + (1-\rho)\gamma^2) - (\rho(1-\epsilon_+)^2 + (1-\rho)\alpha^2) \cdot (\rho\epsilon_- + (1-\rho)\gamma)^2\end{aligned}$$

has a root for $\gamma = \frac{\epsilon_-}{1-\epsilon_+} \cdot \alpha$. [This root was found by an educated guess. First, α was set to equal one and then the $p-q$-formula was applied. The root for general α was hypothesised to be a simple product of this root. This turns out to be true.]

After some algebra we obtain a parabola in γ in standard from:

$$\begin{aligned}&\gamma^2 - \gamma^1 \frac{2\epsilon_- \cdot [\rho(1-\epsilon_+)^2 + \alpha^2(1-\rho)]}{(1-\epsilon_+) \cdot [(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)]}\\ &\quad + \gamma^0 \frac{\epsilon_-^2\alpha \cdot [2\rho(1-\epsilon_+)^2 + \alpha(1-2\rho)]}{(1-\epsilon_+) \cdot [(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)]} \quad .\end{aligned} \tag{B.3}$$

Since the roots of the parabola are equally far away from the axis of symmetry (at γ^* equal to half the fraction following γ^1 in (B.3)), we can compute the second root by subtracting $\alpha\epsilon_+/(1-\epsilon_-)$ from $2\gamma^*$. We find

$$\begin{aligned}&\frac{2\epsilon_- \cdot [\rho(1-\epsilon_+)^2 + \alpha^2(1-\rho)]}{(1-\epsilon_+) \cdot [(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)]} - \alpha\frac{\epsilon_-}{1-\epsilon_+} \cdot \frac{(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)}{(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)}\\ =&\frac{\epsilon_-}{1-\epsilon_+} \cdot \frac{2\rho(1-\epsilon_+)^2 + 2\alpha^2(1-\rho) - \alpha[(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)]}{(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)}\\ =&\frac{\epsilon_-}{1-\epsilon_+} \cdot \frac{2\rho(1-\epsilon_+)^2 - \alpha(1-\epsilon_+)(2\rho-1)}{(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)}\\ =&\frac{\epsilon_- \cdot [2\rho(1-\epsilon_+) + \alpha(1-2\rho)]}{(2\rho-1)(1-\epsilon_+) + 2\alpha(1-\rho)} \quad .\end{aligned}$$

□

Proposition 9. *For the roots of the parabola we find*

$$\gamma_1 := \frac{\epsilon_-}{1-\epsilon_+} \cdot \alpha > \frac{\epsilon_- \cdot [2\rho(1-\epsilon_+) + \alpha(1-2\rho)]}{(2\rho-1)(1-\epsilon_+) + 2\alpha(1-\rho)} =: \gamma_2 \quad .$$

Proof. We find

$$
\begin{aligned}
&\gamma_1 > \gamma_2 \\
\iff &\alpha \cdot [(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)] > 2\rho(1-\alpha) - 2\rho\epsilon_+(2-\alpha) + 2\rho\epsilon_+^2 + \alpha(1-\epsilon_+) \\
\iff &2\alpha\rho(1-\alpha) + 2\alpha^2 - \alpha + \alpha\epsilon_+ - 2\alpha\rho\epsilon_+ > 2\rho(1-\alpha) - 2\rho\epsilon_+(2-\alpha) + 2\rho\epsilon_+^2 + \alpha(1-\epsilon_+) \\
\iff &2\rho(1-\alpha)(\alpha-1) + 2\alpha^2 - 2\alpha + \epsilon_+(\alpha - 2\alpha\rho + 4\rho - 2\alpha\rho - 2\rho\epsilon_+ + \alpha) > 0 \\
\iff &\rho(1-\alpha)(\alpha-1) + \alpha(\alpha-1) + \epsilon_+(\alpha - 2\alpha\rho + 2\rho - \rho\epsilon_+) > 0 \\
\iff &\rho(1-\alpha)\cdot[(\alpha - 1 + \epsilon_+) + \alpha(\alpha - 1 + \epsilon_+) + \epsilon_+\rho(-\alpha + 1 - \epsilon_+) > 0 \\
\iff &(\alpha - 1 + \epsilon_+)\cdot[\rho(1-\alpha) + \alpha - \epsilon_+\rho] > 0 \\
\iff &(\alpha - 1 + \epsilon_+)\cdot[\alpha(1-\rho) + \rho(1-\epsilon_+)] > 0 \ .
\end{aligned}
$$

This holds, since $\alpha > 1 - \epsilon_+$ or equivalently $\alpha + \epsilon_+ - 1 > 0$. □

Theorem 1. *For all $p \in (0,1)$, $q \in (0,p)$, $\rho \in (0,1)$, $\epsilon_+, \epsilon_- \in (0,1)$, $\alpha \in (1-\epsilon_+, 1)$ and $\gamma \in (\epsilon_-, 1)$ the VET fails, if and only if*

$$0 < \gamma_2 \leq \gamma \leq \gamma_1 < 1 \ .$$

Proof. Recall from (B.2) and (B.3) that

$$
\begin{aligned}
&\operatorname{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big) = \operatorname{sign}\Big(T_1T_4 - T_2T_3\Big) \\
&= \operatorname{sign}\Big(\gamma^2 - \gamma^1 \frac{2\epsilon_- \cdot [\rho(1-\epsilon_+)^2 + \alpha^2(1-\rho)]}{(1-\epsilon_+)\cdot[(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)]} + \gamma^0 \frac{\epsilon_-^2\alpha \cdot [2\rho(1-\epsilon_+)^2 + \alpha(1-2\rho)]}{(1-\epsilon_+)\cdot[(1-\epsilon_+)(2\rho-1) + 2\alpha(1-\rho)]}\Big) \ .
\end{aligned}
$$

The last term is a parabola which goes to plus infinity as γ goes to $\pm\infty$ (parabola is open upwards). Hence, the last term is strictly negative, if and only if $\gamma_2 < \gamma < \gamma_1$. The posterior probabilities are equal, if and only if $\gamma \in \{\gamma_1, \gamma_2\}$. The VET holds, if and only if $\gamma \notin [\gamma_2, \gamma_1]$. □

Proposition 10. *For all $\alpha, \gamma, \rho \in (0,1)$, $\epsilon_+ \in (0, \alpha)$, $\epsilon_- \in (0, \gamma)$ and all $N \geq 3$ positive reports*

- *The VET fails for $\gamma \leq \frac{\epsilon_-}{1-\epsilon_+}\alpha$ where γ is close to $\frac{\epsilon_-}{1-\epsilon_+}$.*
- *The VET holds for all $\gamma > \frac{\epsilon_-}{1-\epsilon_+}\alpha$.*

Proof. From (B.1) we have that $\text{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big)$ is equal to the sign of

$$T_1T_4 - T_2T_3 =$$
$$(\rho(1-\epsilon_+) + (1-\rho)\alpha)^N \cdot (\rho\epsilon_-^N + (1-\rho)\gamma^N) - (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) \cdot (\rho\epsilon_- + (1-\rho)\gamma)^N \ .$$

For $\gamma = \frac{\epsilon_-}{1-\epsilon_+} \cdot \alpha$ this becomes

$$\begin{aligned}
&(\rho(1-\epsilon_+) + (1-\rho)\alpha)^N \cdot (\rho\epsilon_-^N + (1-\rho)\frac{\epsilon_-^N}{(1-\epsilon_+)^N} \cdot \alpha^N) \\
&\quad - (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) \cdot (\rho\epsilon_- + (1-\rho)\frac{\epsilon_-}{1-\epsilon_+} \cdot \alpha)^N \\
=&\frac{\epsilon_-^N}{(1-\epsilon_+)^N} \cdot [(\rho(1-\epsilon_+) + (1-\rho)\alpha)^N \cdot (\rho(1-\epsilon_+)^N + (1-\rho) \cdot \alpha^N) \\
&\quad - (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) \cdot (\rho(1-\epsilon_+) + (1-\rho) \cdot \alpha)^N] \\
=&\ 0 \ .
\end{aligned}$$

For $\gamma \neq \frac{\epsilon_-}{1-\epsilon_+}\alpha$ we define Δ implicitly by $\gamma = \frac{\epsilon_-}{1-\epsilon_+}\alpha \cdot (1+\Delta)$. We are interested in the area where Δ is very close to zero and hence do a Taylor expansion in the following ignoring terms of order Δ^2 and higher of

$$\begin{aligned}
&(\rho(1-\epsilon_+) + (1-\rho)\alpha)^N \cdot (\rho\epsilon_-^N + (1-\rho)\frac{\epsilon_-^N}{(1-\epsilon_+)^N}\alpha^N(1+\Delta)^N) \\
&\quad - (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) \cdot (\rho\epsilon_- + (1-\rho)\frac{\epsilon_-}{1-\epsilon_+}\alpha(1+\Delta))^N \\
=&\frac{\epsilon_-^N}{(1-\epsilon_+)^N} \cdot [(\rho(1-\epsilon_+) + (1-\rho)\alpha)^N \cdot (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N(1+\Delta)^N) \\
&\quad - (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) \cdot (\rho(1-\epsilon_+) + (1-\rho)\alpha(1+\Delta))^N] \\
=&\frac{\epsilon_-^N}{(1-\epsilon_+)^N} \cdot [(\rho(1-\epsilon_+) + (1-\rho)\alpha)^N \cdot (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N(1+\Delta)^N) \\
&\quad - (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) \cdot (\alpha - \rho(\epsilon_+ + \alpha - 1) + \alpha\Delta(1-\rho))^N] \ .
\end{aligned}$$

The constant term, Δ^0 is, modulo the leading strictly positive factor

$$\begin{aligned}
&(\rho(1-\epsilon_+) + (1-\rho)\alpha)^N \cdot (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) - [(\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) \cdot (\alpha - \rho(\epsilon_+ + \alpha - 1))^N] \\
=&(\alpha - \alpha\rho + \rho - \rho\epsilon_+)^N \cdot (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N) - [(\alpha - \alpha\rho + \rho - \rho\epsilon_+)^N \cdot (\rho(1-\epsilon_+)^N + (1-\rho)\alpha^N)]
\end{aligned}$$

$=0$.

For the term of order one, Δ^1, we find, modulo the leading strictly positive factor:

$$
\begin{aligned}
&(\rho(1-\epsilon_+)+(1-\rho)\alpha)^N\cdot(1-\rho)\alpha^N N\Delta-(\rho(1-\epsilon_+)^N+(1-\rho)\alpha^N)\cdot(\alpha-\rho(\epsilon_++\alpha-1))^{N-1}(1-\rho)\alpha N\Delta\\
=&(\rho(1-\epsilon_+)+(1-\rho)\alpha)^{N-1}(1-\rho)\alpha N\Delta\cdot[(\rho(1-\epsilon_+)+(1-\rho)\alpha)\alpha^{N-1}-\rho(1-\epsilon_+)^N-(1-\rho)\alpha^N]\\
=&(\rho(1-\epsilon_+)+(1-\rho)\alpha)^{N-1}(1-\rho)\alpha N\Delta\cdot[\rho(1-\epsilon_+)\alpha^{N-1}-\rho(1-\epsilon_+)^N]\\
=&(\rho(1-\epsilon_+)+(1-\rho)\alpha)^{N-1}(1-\rho)\alpha N\Delta\rho(1-\epsilon_+)\cdot[\alpha^{N-1}-(1-\epsilon_+)^{N-1}]\ ,
\end{aligned}
$$

Since $\alpha > 1-\epsilon_+$, the sign of this expression is equal to the sign of Δ.

So, for $\gamma > \frac{\epsilon_-}{1-\epsilon_+}\alpha$ but close to $\frac{\epsilon_-}{1-\epsilon_+}\alpha$ (that is $\Delta > 0$) the VET holds and for $\gamma < \frac{\epsilon_-}{1-\epsilon_+}\alpha$ but close to $\frac{\epsilon_-}{1-\epsilon_+}\alpha$ (that is $\Delta < 0$) the VET fails.

For the higher orders of Δ we find

$$
\begin{aligned}
&(\rho(1-\epsilon_+)+(1-\rho)\alpha)^N\alpha^N(1-\rho)\cdot\sum_{k=2}^{N}\binom{N}{k}\Delta^k\\
-&[\rho(1-\epsilon_+)^N+(1-\rho)\alpha^N]\cdot\sum_{k=2}^{N}\binom{N}{k}(\rho(1-\epsilon_+)+(1-\rho)\alpha)^{N-k}\alpha^k(1-\rho)^k\Delta^k\\
=&(1-\rho)\cdot\sum_{k=2}^{N}\binom{N}{k}\Delta^k(\rho(1-\epsilon_+)+(1-\rho)\alpha)^{N-k}\alpha^k\cdot\\
&\qquad[(\rho(1-\epsilon_+)+(1-\rho)\alpha)^k\alpha^{N-k}-(\rho(1-\epsilon_+)^N+(1-\rho)\alpha^N)(1-\rho)^{k-1}]\ .
\end{aligned}
$$

For $\Delta > 0$, this strictly positive, if the expression in square brackets is positive. For this expression we find, since $\alpha > 1-\epsilon_+$

$$
\begin{aligned}
&(\rho(1-\epsilon_+)+(1-\rho)\alpha)^k\alpha^{N-k}-(\rho(1-\epsilon_+)^N+(1-\rho)\alpha^N)(1-\rho)^{k-1}\\
>&(\rho(1-\epsilon_+)+(1-\rho)\alpha)^k\alpha^{N-k}-(\rho(\alpha)^N+(1-\rho)\alpha^N)(1-\rho)^{k-1}\\
=&\alpha^{N-k}\cdot[(\rho(1-\epsilon_+)+(1-\rho)\alpha)^k-\alpha^k(1-\rho)^{k-1}]\ .
\end{aligned}
$$

$$
0 < \rho^2(1-\epsilon_+)^2
$$

$$
\begin{aligned}
&= \rho^2(1-\epsilon_+)^2 + \alpha^2\rho(1-\rho) - \alpha^2\rho + \alpha^2\rho^2 \\
&\leq \rho^2(1-\epsilon_+)^2 + 2\rho(1-\epsilon_+)(1-\rho)\alpha - \alpha^2\rho + \alpha^2\rho^2 \\
&= \rho^2(1-\epsilon_+)^2 + 2\rho(1-\epsilon_+)(1-\rho)\alpha + \alpha^2 - 2\alpha^2\rho + \alpha^2\rho^2 - \alpha^2 + \alpha^2\rho \\
&= (\rho(1-\epsilon_+) + (1-\rho)\alpha)^2 - \alpha^2(1-\rho)^1 \\
&\leq (\rho(1-\epsilon_+) + (1-\rho)\alpha)^2 - \alpha^2(1-\rho)^{2(k-1)/k} \quad .
\end{aligned}
$$

Having established that this difference is strictly positive, we can raise the minuend and the subtrahend to the power $k/2$ and retain a strictly positive difference. This completes the proof. □

Proposition 11. *$\alpha > 1 - \epsilon_+$ entails that $\gamma_2 < \epsilon_-$*

Proof. It suffices to show that

$$
\gamma_2 = \frac{\epsilon_- \cdot [2\rho(1-\epsilon_+) + \alpha(1-2\rho)]}{(2\rho-1)(1-\epsilon_+) + 2\alpha(1-\rho)} < \epsilon_- \quad .
$$

This is equivalent to $(1-\epsilon_+) + \alpha(1-2\rho) < 2\alpha(1-\rho)$. Which is in turn equivalent to $1 - \epsilon_+ < \alpha$ which holds in our model. □

B.1.2 Scenario 2

We being by calculating likelihoods for N positive reports:

$$
\begin{aligned}
P_1(E|Hyp) &= \rho(p(1-\epsilon) + (1-p)\epsilon)^N + (1-\rho)(p\alpha + (1-p)\gamma)^N \\
P_1(E|\overline{Hyp}) &= \rho(q(1-\epsilon) + (1-q)\epsilon)^N + (1-\rho)(q\alpha + (1-q)\gamma)^N \\
P(E|Hyp) &= \rho(p(1-\epsilon) + (1-p)\epsilon) + (1-\rho)(p\alpha + (1-p)\gamma) \\
P(E|\overline{Hyp}) &= \rho(q(1-\epsilon) + (1-q)\epsilon) + (1-\rho)(q\alpha + (1-q)\gamma) \quad .
\end{aligned}
$$

By Lemma 1 we have

$$
\text{sign}(P(Hyp|E) - P_1(Hyp|E)) = \text{sign}\Big(P_1(E|\overline{Hyp}) \cdot P(E|Hyp) - P(E|\overline{Hyp}) \cdot P_1(E|Hyp)\Big) \quad .
$$

The VET holds, if and only if the sign of this expression is negative. We shall only consider $N = 2$.

We obtain a polynomial of degree three in the variable γ which we can solve using standard techniques. The coefficients are sorted in powers of α, they hence cannot be simplified. The coefficients are

$$
\begin{aligned}
g_3 :=&(1-\rho)(1-q)^2(1-\rho)(1-p) - (1-\rho)(1-p)^2(1-\rho)(1-q) \\
=&(1-\rho)^2(1-p)(1-q)(p-q) \\
g_2 :=&(1-\rho)(1-q)^2(\rho(p(1-\epsilon)+(1-p)\epsilon)) - (1-\rho)(1-p)^2(\rho(q(1-\epsilon)+(1-q)\epsilon)) \\
&+ (1-q)^2(1-r)^2 p\alpha - (1-p)^2(1-\rho)^2 q\alpha \\
&+ 2(1-q)q\alpha(1-\rho^2)(1-p) - 2(1-p)p\alpha(1-\rho)^2(1-q) \\
=&\rho(1-\rho)\cdot[p-q+qp(q-p) - 2pq\epsilon(q-p) + \epsilon(q-p)(q+p)] \\
&+ (1-\rho)^2\alpha(p-q)(1-pq) \\
&+ 2(1-p)(1-q)\alpha(1-\rho)^2(q-p) \\
=&\rho(1-\rho)\cdot[p-q+qp(q-p) - 2pq\epsilon(q-p) + \epsilon(q-p)(q+p)] \\
&+ (1-\rho)^2\alpha(p-q)(-1-3pq+2p+2q) \\
g_1 :=&[\rho(q(1-\epsilon)+(1-q)\epsilon)^2 + (1-\rho)(q\alpha)^2]\cdot(1-\rho)(1-p) \\
&+ 2(1-\rho)q\alpha(1-q)\cdot[\rho(p(1-\epsilon)+(1-p)\epsilon) + (1-\rho)p\alpha] \\
&- [\rho(p(1-\epsilon)+(1-p)\epsilon)^2 + (1-\rho)(p\alpha)^2]\cdot(1-\rho)(1-q) \\
&- 2(1-\rho)p\alpha(1-p)\cdot[\rho(q(1-\epsilon)+(1-q)\epsilon) + (1-\rho)q\alpha] \\
=&(1-\rho)^2\alpha^2(q-p)(q+p-qp) + 2(1-\rho)^2\alpha^2 pq(p-q) \\
&+ [\rho(q(1-\epsilon)+(1-q)\epsilon)^2]\cdot(1-\rho)(1-p) \\
&+ 2(1-\rho)q\alpha(1-q)\cdot[\rho(p(1-\epsilon)+(1-p)\epsilon)] \\
&- [\rho(p(1-\epsilon)+(1-p)\epsilon)^2]\cdot(1-\rho)(1-q) \\
&- 2(1-\rho)p\alpha(1-p)\cdot[\rho(q(1-\epsilon)+(1-q)\epsilon)] \\
=&\alpha^2(1-\rho)^2(q-p)(q+p-3qp) \\
&+ \alpha 2\rho(1-\rho)(1-\epsilon)pq(p-q) + \alpha 2\rho(1-\rho)\epsilon(1-p)(1-q)(q-p)
\end{aligned}
$$

$$
\begin{aligned}
&- \rho(1-\rho)\cdot[(q-p)(q+p) + 4\epsilon^2(q^2-p^2) + 2\epsilon(q-p) - 4\epsilon(q-p)(q+p) \\
&\quad - 4\epsilon^2(q-p) + pq(p-q) + \epsilon^2(q-p) + 4\epsilon^2 pq(p-q) + 4\epsilon pq(q-p)] \\
g_0 :=&[\rho(q(1-\epsilon) + (1-q)\epsilon)^2 + (1-\rho)(q\alpha)^2]\cdot[\rho(p(1-\epsilon) + (1-p)\epsilon) + (1-\rho)p\alpha] \\
&- [\rho(p(1-\epsilon) + (1-p)\epsilon)^2 + (1-\rho)(p\alpha)^2]\cdot[\rho(q(1-\epsilon) + (1-q)\epsilon) + (1-\rho)q\alpha] \\
=&\rho^2(q+\epsilon-2q\epsilon)\cdot(p+\epsilon-2p\epsilon)\cdot[q-p+2\epsilon(p-q)] \\
&+ \alpha\rho(1-\rho)\cdot[pq(q-p) + 4\epsilon pq(p-q) + \epsilon^2(p-q) + 4\epsilon^2 pq(q-p)] \\
&+ \alpha^2\rho(1-\rho)[pq(q-p) + \epsilon(p+q)(q-p) + 2\epsilon pq(p-q)] + \alpha^3(1-\rho)^2 pq(q-p) \ .
\end{aligned}
$$

We obtain the general – utterly intractable – formula

$$
\operatorname{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big) = \operatorname{sign}\Big(g_3\gamma^3 + g_2\gamma^2 + g_1\gamma + g_0\Big) \ . \tag{B.4}
$$

B.1.3 Scenario 3

Theorem 2. *For all $p \in (0,1)$, $q \in (0,p)$, $\rho \in (0,1)$, $\epsilon_+, \epsilon_- \in (0,1)$, $\alpha \in (\epsilon_+, 1)$, $\gamma \in (\epsilon_-, 1)$ and all $N \geq 2$, if $\gamma = \alpha \cdot \frac{\epsilon_-}{1-\epsilon_+}$, then*

$$
P(Hyp|E) = P_1(Hyp|E) \ .
$$

Proof. We have

$$
\begin{aligned}
P_1(E|Hyp) =&[\rho(p(1-\epsilon_+) + (1-p)\epsilon_-) + (1-\rho)(p\alpha + (1-p)\gamma)]^N \\
P_1(E|\overline{Hyp}) =&[\rho(q(1-\epsilon_+) + (1-q)\epsilon_-) + (1-\rho)(q\alpha + (1-q)\gamma)]^N \\
P(E|Hyp) =&\rho(p(1-\epsilon_+) + (1-p)\epsilon_-)^N + (1-\rho)(p\alpha + (1-p)\gamma)^N \\
P(E|\overline{Hyp}) =&\rho(q(1-\epsilon_+) + (1-q)\epsilon_-)^N + (1-\rho)(q\alpha + (1-q)\gamma)^N \ .
\end{aligned}
$$

By Lemma 1 we have

$$
\begin{aligned}
&\operatorname{sign}\Big(P(Hyp|E) - P_1(Hyp|E)\Big) \\
&\quad = \operatorname{sign}\Big(P_1(E|\overline{Hyp})\cdot P(E|Hyp) - P(E|\overline{Hyp})\cdot P_1(E|Hyp)\Big) \ .
\end{aligned}
$$

The VET holds, if and only if the sign of this expression is negative.

Again, it is much too long to be tractably solved by us. While much shorter than in Scenario 2, this polynomial only spans a handful of pages.

Luckily, we can guess a solution:[29] $\gamma = \alpha \cdot \frac{\epsilon_-}{1-\epsilon_+}$. Plugging this in, we obtain

$$\begin{aligned}
P_1(E|Hyp) =& [\rho(p(1-\epsilon_+) + (1-p)\epsilon_-) + (1-\rho)(p\alpha + (1-p)\alpha\frac{\epsilon_-}{1-\epsilon_+})]^N \\
=& [\rho(p(1-\epsilon_+) + (1-p)\epsilon_-) + (1-\rho)\frac{\alpha}{(1-\epsilon_+)}(p(1-\epsilon_+) + (1-p)\epsilon_-)]^N \\
=& (p(1-\epsilon_+) + (1-p)\epsilon_-)^N[\rho + (1-\rho)\frac{\alpha}{(1-\epsilon_+)}]^N \\
P_1(E|\overline{Hyp}) =& (q(1-\epsilon_+) + (1-q)\epsilon_-)^N[\rho + (1-\rho)\frac{\alpha}{(1-\epsilon_+)}]^N \\
P(E|Hyp) =& \rho(p(1-\epsilon_+) + (1-p)\epsilon_-)^N + (1-\rho)(p\alpha + (1-p)\alpha\frac{\epsilon_-}{1-\epsilon_+}))^N \\
=& \rho(p(1-\epsilon_+) + (1-p)\epsilon_-)^N + (1-\rho)\frac{\alpha^N}{(1-\epsilon_+)^N}(p(1-\epsilon_+)\alpha + (1-p)\epsilon_-))^N \\
=& [\rho + (1-\rho)\frac{\alpha^N}{(1-\epsilon_+)^N}] \cdot [(p(1-\epsilon_+) + (1-p)\epsilon_-)^N] \\
P(E|\overline{Hyp}) =& [\rho + (1-\rho)\frac{\alpha^N}{(1-\epsilon_+)^N}] \cdot [(q(1-\epsilon_+) + (1-q)\epsilon_-)^N] \ .
\end{aligned}$$

It follows immediately that for $\gamma = \alpha\epsilon_-/(1-\epsilon_+)$

$$P(Hyp|E) = P_1(Hyp|E) \ .$$

□

B.2 Graphical Exploration of Parameter Spaces in Our Model

We here present instances of our graphical exploration of parameter spaces for $N \geq 3$ reports.

[29] The solution was not discovered by pure chance. First, we created a number of plots as described under Scenario 2 for $\epsilon_+ = \epsilon_-$. This time, we found parameter values where the VET fails, for $N = 2$ (see Figure 7). For these parameter values we plotted the curves in the ϵ, γ-plane where both posteriors are equal. One of the curves looked eerily familiar. Verily, upon feeding the critical parameters to Octave, we obtained $\gamma = \frac{-39\epsilon}{40\epsilon-40}$. As it happened, $\alpha = 39/40$ we thus rediscovered $\gamma = \alpha\epsilon/(1-\epsilon)$. Jackpot!

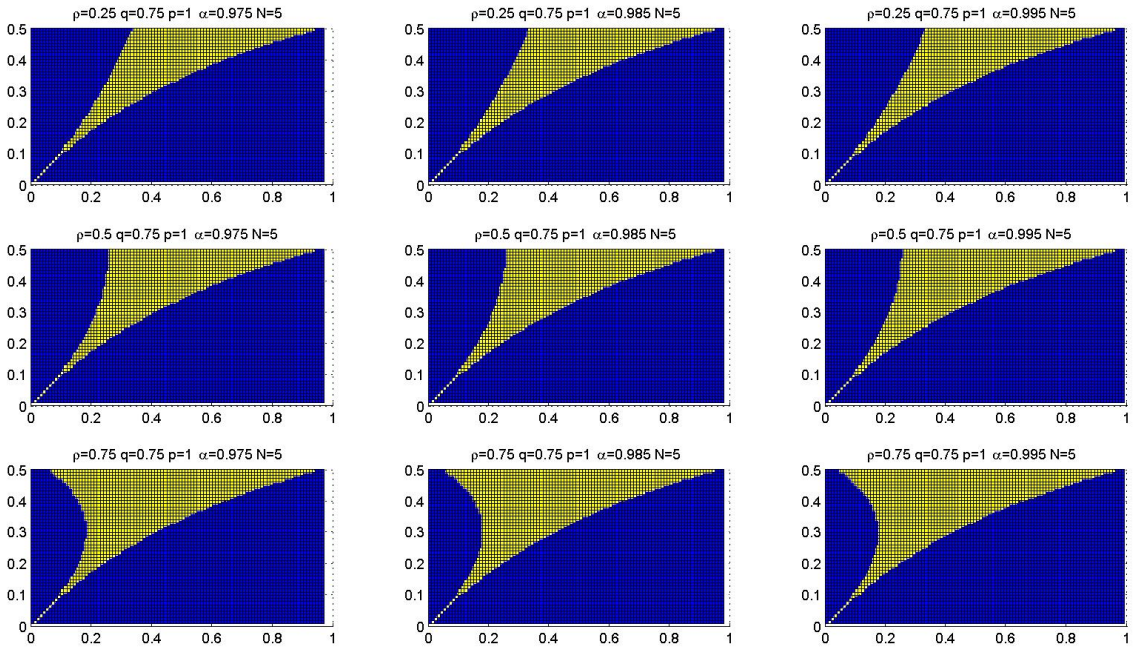

Figure 16: **Scenario 1: The $\gamma - \epsilon$-plane for varying ρ, α and fixed $\mathbf{p = 0.9, q = 0.5, N = 5}$.** Yellow indicates the area of VET failure. Within a column ρ varies: top $\rho = 0.25$, middle $\rho = 0.5$, bottom $\rho = 0.75$. Within a row α varies: left $\alpha = 0.975$, middle $\alpha = 0.985$, right $\alpha = 0.995$.

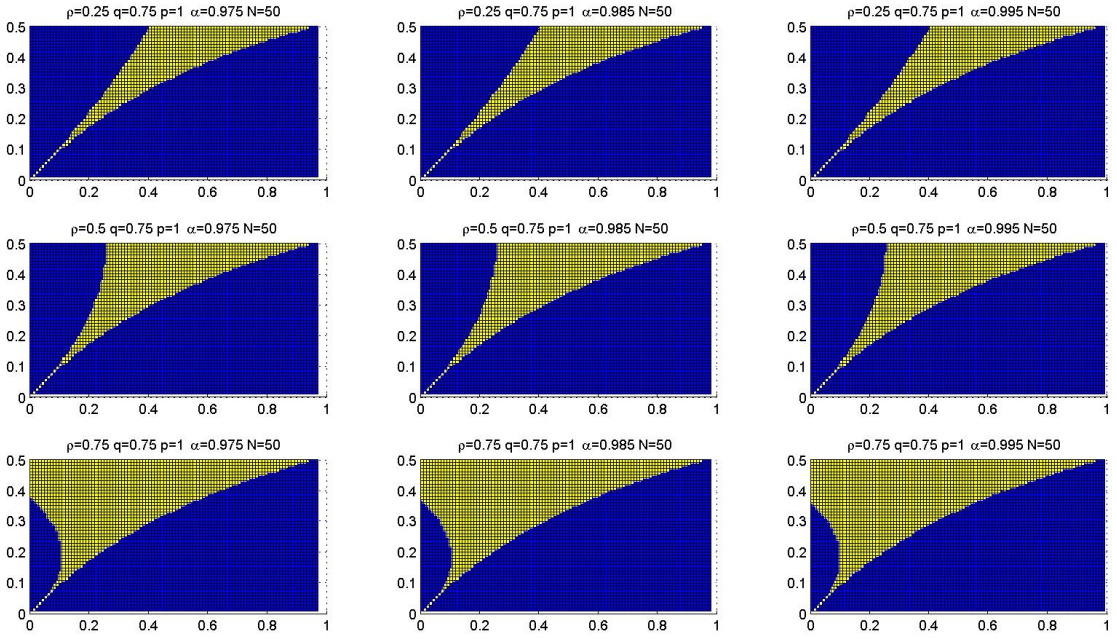

Figure 17: **Scenario 1: The $\gamma - \epsilon$-plane for varying ρ, α and fixed $\mathbf{p = 0.9, q = 0.5, N = 50}$.** Yellow indicates the area of VET failure. Within a column ρ varies: top $\rho = 0.25$, middle $\rho = 0.5$, bottom $\rho = 0.75$. Within a row α varies: left $\alpha = 0.975$, middle $\alpha = 0.985$, right $\alpha = 0.995$.

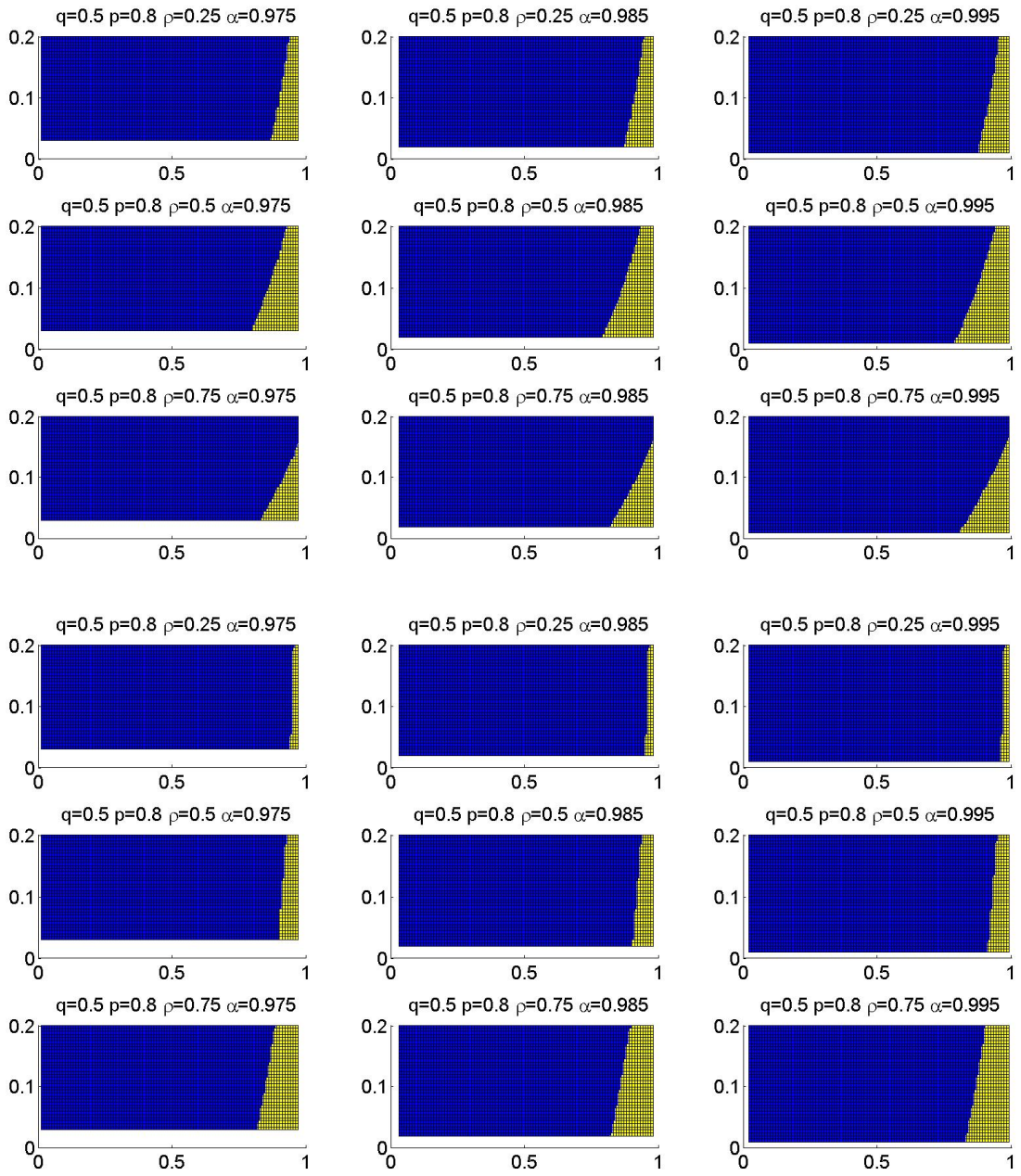

Figure 18: **Scenario 2: The $\gamma - \epsilon$-plane for varying ρ, α, N and fixed p $= 0.9$, q $= 0.5$.** Yellow indicates the area of VET failure which is only observed for large γ. In the top 3x3 set $N = 2$, in the bottom set $N = 10$. Within a column of a set ρ varies: top $\rho = 0.25$, middle $\rho = 0.5$, bottom $\rho = 0.75$. Within a row α varies: left $\alpha = 0.975$, middle $\alpha = 0.985$, right $\alpha = 0.995$.

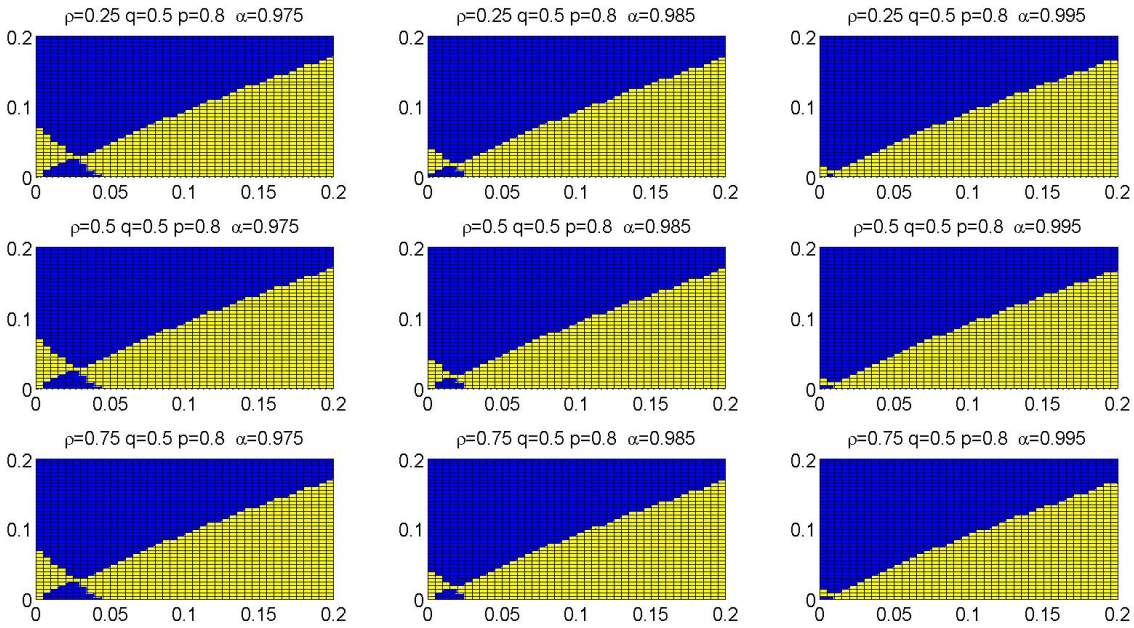

Figure 19: **Scenario 3: The $\gamma - \epsilon$-plane for varying ρ, α and fixed $\mathbf{p = 0.8, q = 0.5, N = 5}$.** Yellow indicates the area of VET failure. Within a column of a set ρ varies: top $\rho = 0.25$, middle $\rho = 0.5$, bottom $\rho = 0.75$. Within a row α varies: left $\alpha = 0.975$, middle $\alpha = 0.985$, right $\alpha = 0.995$.

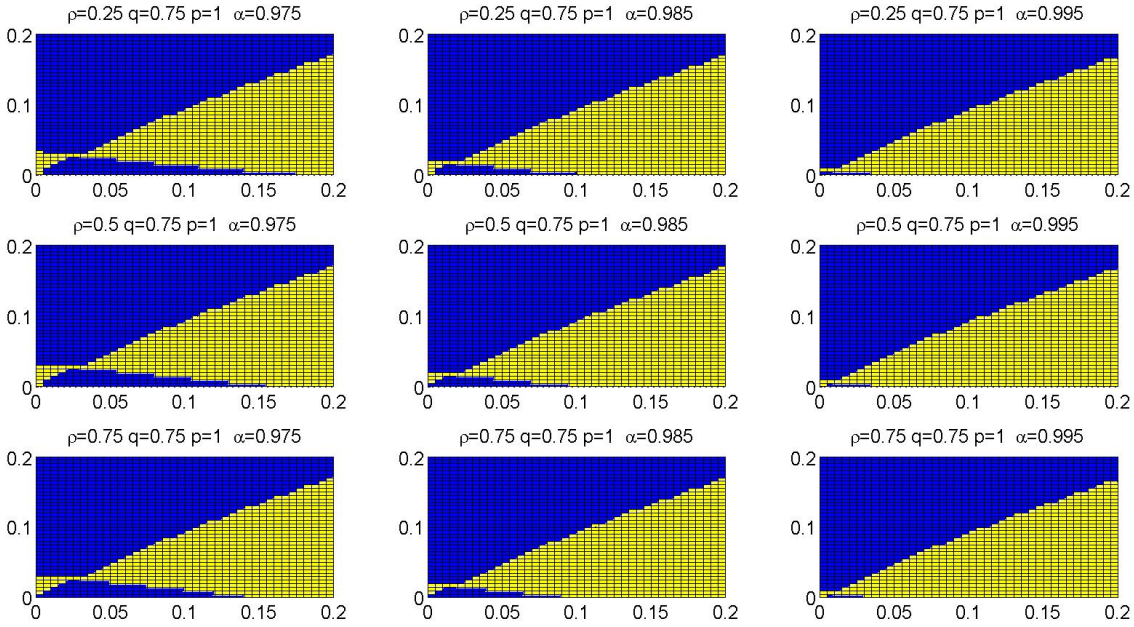

Figure 20: **Scenario 3: The $\gamma - \epsilon$-plane for varying ρ, α and fixed $\mathbf{p = 1, q = 0.75, N = 50}$.** Yellow indicates the area of VET failure. Within a column of a set ρ varies: top $\rho = 0.25$, middle $\rho = 0.5$, bottom $\rho = 0.75$. Within a row α varies: left $\alpha = 0.975$, middle $\alpha = 0.985$, right $\alpha = 0.995$.

Appendix C Online Appendix

C.1 Further Discussion of Bovens and Hartmann's Scenario 2

In Figure 21, the VET fails in the $p - q$-plane, with a fixed at 0.5, for the points underneath the curves $\rho = 0.1$ (blue curve), $\rho = 0.5$ (green curve) and $\rho = 0.9$ (red curve).[30] For infinitely many reports the VET fails, if and only if $p + q < 0.5$ (purple); see Proposition 3. That is the probability of observing a positive report, no matter whether the hypothesis holds or not, is below 0.5. In such conditions, receiving consistent positive reports from the same instrument, under the assumption that, if it is unreliable, it is a proper randomiser ($a = 0.5$), boosts the hypothesis less than receiving just one report.

In Figure 22, the VET fails in the $p - q$-plane with ρ fixed at 0.5 for the points underneath the curves $a = 0.1$ (blue curve), $a = 0.5$ (green curve) and $a = 0.9$ (red curve).

In this second scenario the structure of the body of evidence (more vs. less varied) interacts both with ρ and a as well as with the "strength" of the indicator. A strong indicator is a consequence-variable with high p and low q: this means that the consequence is strongly correlated with the hypothesis of interest and has also high discriminatory power for such hypothesis with respect to its alternatives. A weak indicator ($p \approx q$) may be so in two different ways: the consequence tends to hold no matter whether HYP is true or not (both p and q are high); or, on the contrary, it rarely occurs (both p and q are low). In the former case, one expects a reliable instrument to deliver more positive reports than not (somewhat independently of the truth of HYP), whereas in the second case, one expects the reverse. This fully explains VET failure, in conjunction with the value assumed by a. Analogous considerations hold for the following third scenario.

In Figure 22 (Scenario 2), the area relevant for our analysis is below the $p = q$ line (cyan). The bottom graphs display the area of VET failure for 50 items of confirming evidence, for $a = 0.1$ and $a = 0.9$ respectively. We notice that the area of VET failure grows with an increasing number of items of incoming evidence. Matters are more complicated in Scenario 3

[30] By the assumption that the testable consequences of the hypothesis are positively correlated with the hypothesis we have $p > q$.

(a snapshot is provided in Figure 4), where we see an interaction between a, ρ and p and q with the number of reports; however belief dynamics follow the same lines of reasoning provided for $N = 2$ scenarios in the previous section.

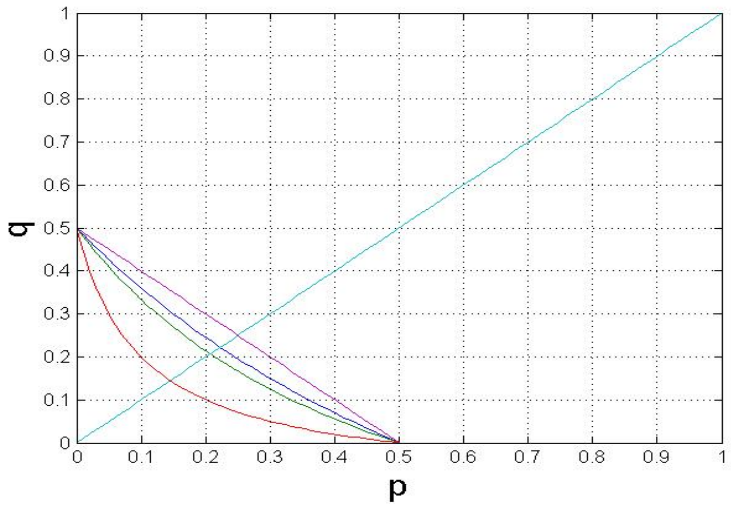

Figure 21: **Figure 4.8 from (Bovens and Hartmann 2003, p. 100), Scenario 2.** The VET fails in the $p - q$-plane for $a = 0.5$ for the points underneath the curves $\rho = 0.1$ (blue), $\rho = 0.5$ (green) and $\rho = 0.9$ (red). $p > q$ is also assumed in the Bovens and Hartmann model (cyan). In the limit, the VET fails if and only if $p + q < 0.5$ (purple). We again notice that the area of VET failure grows.

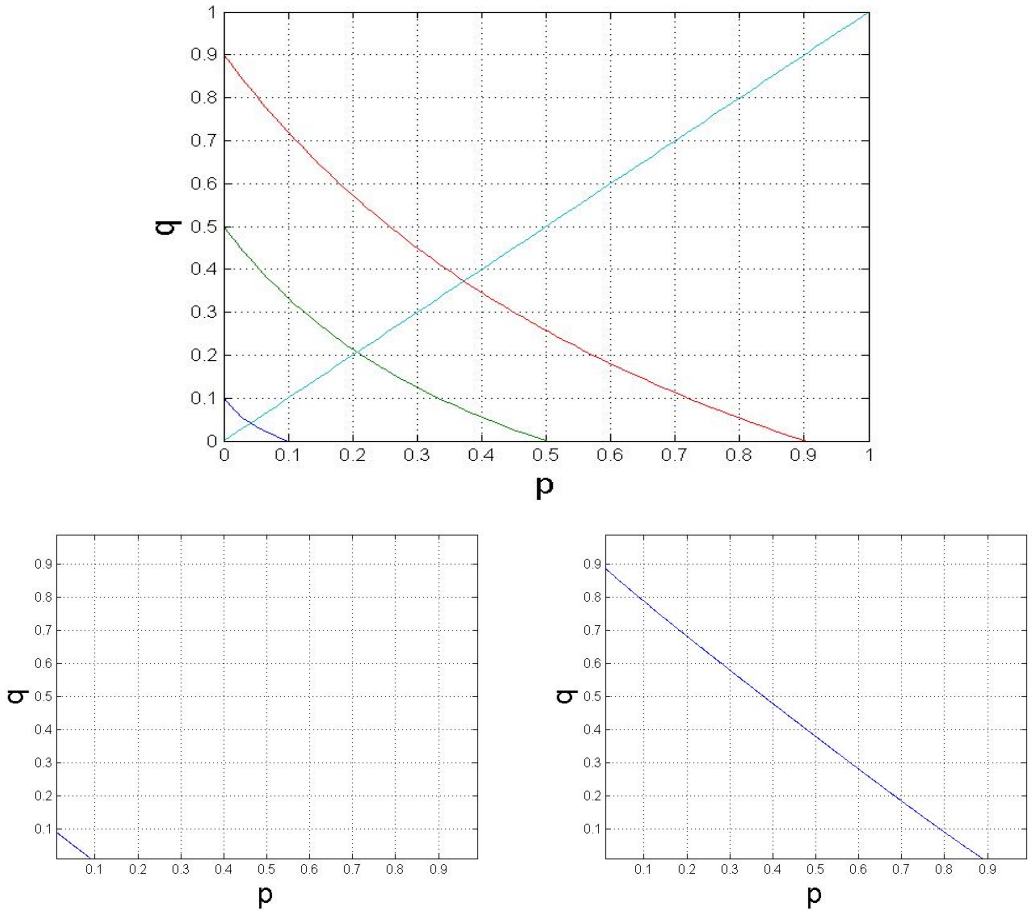

Figure 22: **Figure 4.9 from (Bovens and Hartmann 2003, p. 101) on top, Scenario 2.** The VET fails in the $p - q$-plane for $\rho = 0.5$ for the points underneath the curves $a = 0.1$ (blue), $a = 0.5$ (green) and $a = 0.9$ (red). $p > q$ is also assumed in the Bovens and Hartmann model (cyan). The bottom graphs display the area of VET failure for $a = 0.1$ and $a = 0.9$, respectively, for 50 items of confirming evidence. We again notice that the area of VET failure grows.