Title

Shared Evidence: It all depends…

Permalink

https://escholarship.org/uc/item/5008d3h5

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 41(0)

Authors

Pilditch, Toby D.

Hahn, Ulrike

Lagnado, David

Publication Date

2019

Peer reviewed

Shared Evidence: It all depends...

Toby D. Pilditch[1,2], Ulrike Hahn[3], and David Lagnado[1]

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK
[2]University of Oxford, School of Geography and the Environment, South Parks Road, Oxford, OX1 3QY, UK
[3]Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK

Abstract

When reasoning about evidence, we must carefully consider the impact of different structures. For instance, if in the process of evaluating multiple reports, we find they rely on the same, *shared* evidence, then the support proffered by those reports is dependent on that evidence. Critically, normative accounts suggest that such a dependency results in redundant information across reports (reducing evidential support), relative to reports based on distinct items of evidence. In the present work we disentangle the structural and observation-based indicators of this form of dependency. In so doing, we present novel findings that lay reasoners are not only insensitive to shared evidence structures when updating their beliefs, but also that reasoners do not necessarily prefer more diverse sources of evidence. Finally, we replicate prior effects in reasoning under uncertainty, including conservative sequential updating, and difficulty in integrating contradictory reports.

Keywords: evidential reasoning; probabilistic reasoning; dependence; Bayesian Networks; belief updating

Introduction

Over the course of an investigation, you are faced with the weighing up of contradicting reports. Two of your investigators confirm the hypothesis, whilst two disconfirm it. How do you discern which pair may carry more (evidential) weight? One important aspect is what evidence those investigators are relying upon. For instance, if your two confirming investigators are relying on the *same* piece of evidence to inform their reports, whilst the two disconfirming investigators are relying on separate, independent pieces of evidence, then, *ceterus paribus*, the standard intuition is to side with the disconfimers.

This example highlights the traditional understanding of one form of dependency in evidential reasoning. Specifically, the notion of “shared” evidence (Schum & Martin, 1982; Schum, 1994), which is considered to be inferior to reports based on distinct (separate) evidence, i.e. dependence as a form of redundancy (Hogarth, 1989; Schum & Martin, 1982; Soll, 1999).

How such information *should* be integrated is important to a number of areas, from everyday reasoning to investigative domains such as medicine (Eddy, 1982), law (Faigman & Baglioni Jr, 1988; Fenton & Neil, 2012, Fenton, Neil & Lagnado, 2013, Harris & Hahn, 2009; Lagnado, 2011; Pennington & Hastie, 1986; Schum, 1994), risk analysis (Fenton & Neil, 2012), and to the intelligence community (Heuer, 1999). Consequently, failures to account for such dependencies between evidence items – although easing computation (Pearl, 1988; Schum, 1994) – can lead to deleterious overweighting of the support provided by such evidence (e.g., naïve Bayes in medicine – where evidence is assumed to always be independent; Koller & Friedman, 2009; Kononenko, 1993).

The notion of shared-evidence as a form of dependence fits with the correlation-based conceptualisation of dependencies as a form of redundancy in prediction errors (e.g. Soll, 1999). More precisely, when two sources are using the same evidence to inform their reports, vs the same two sources using two different items of evidence, the former case results in an “overlap” of information provided (Schum, 1994). It thus becomes more likely that reports in the former case rely on the *same* information, and the pair of reports therefore carry some redundant information. As a consequence, such correlated reports provide a lesser degree of support for the hypothesis being informed upon.

In the present work, we seek to provide an empirical baseline for lay reasoners judgments regarding this form of dependence. We not only investigate whether belief-updating is in line with the shared-evidence-as-inferior hypothesis proposed in formal work, but whether lay reasoners seek more diverse evidence in their search preferences.

Formalising reasoning about shared evidence

To illustrate what is meant by shared evidence, Fig. 1 below presents a directed acyclic graph (DAG) of an example case. Here there is a hypothesis under investigation (H), three pieces of evidence that inform that hypothesis (E1-3), and four sources (or witnesses) who in turn *report* on said evidence (R1-4). Crucially, the evidence itself remains unobserved, so we are instead trying to infer diagnostically about H (via E1-3) from the reports provided by R1-4, and notably how to judge R1 and R2 (who rely on the same evidence, E1), versus R3 and R4 (who rely on separate evidence, E2 and E3 respectively).

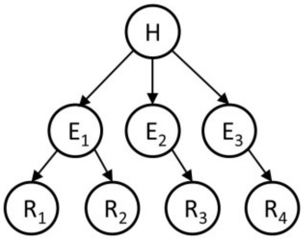

Figure 1. Graphical representation of a hypothesis (H), evidence items that inform upon it (E1-E3), and sources informing their reports upon said evidence (R1-R4).

To understand how reasoners *should* update their beliefs given these observed reports, we use a Bayesian Network (BN) formalism, wherein a DAG is supplemented by conditional probabilities and the use of Bayes theorem, so as to make optimal (i.e. inaccuracy minimization) inferences (Pearl, 1988; 2009). This computational framework for reasoning under uncertainty has been used effectively to model (and shed light on human inferences by comparison) direct dependencies between sources (Pilditch, Hahn & Lagnado, 2018), and dependencies as shared-backgrounds among sources (Madsen, Hahn & Pilditch, 2018), and integration across sources of differing reliabilities (Phillips, Hahn, & Pilditch, 2018).

If we assume, in the above example, that all evidence items are equally diagnostic[1], and all sources are equally reliable[2], then the sole difference-maker between sources is the structural difference entailed by the shared evidence (E1). To best illustrate the impact of shared evidence, we first consider the point at which we have only observed a confirmatory report from R1. Via conditionalization, E1 is now already more likely to be confirmatory than E2 and E3. Given this, if we are to decide whether we want to see a report from R2 (who also relies on E1), or a report from R3, a confirming report from the latter provides *more potential information* regarding H, given that P(E1) – which is already more probable, given the report from R1 – increases less given R2 than P(E2) does, given R3.

Present research We seek to empirically test the degree to which lay reasoners are sensitive to the impact of shared evidence structures on belief updating. More precisely, we use the above formalism to provide an empirical baseline for lay reasoning regarding such dependencies, and notably whether participant probability estimates fit with normative predictions of dependence inferiority. Additionally, we explore two research questions that the formalism allows us to investigate, via the separation of *structural* dependencies from dependencies inferred from (correlated) observations:

First, how do reasoners deal with contradiction across a shared evidence item (as opposed to contradiction across different evidence items)? Recent research that exploits the capacity to tease apart the structural form of a dependency from the dependency inferred from (correlated) observation – as possible in the present work – exposes lay reasoner difficulties in accurately updating (both qualitatively and quantitatively) when an observed contradiction occurs *across* a structural dependency (i.e. information is directly shared from one source to another equally reliable source, yet those sources then disagree; Pilditch, Hahn, & Lagnado, 2018). We predict the same difficulty here.

Second, the present work allows for the investigation into evidence diversity preferences. The computational framework underpinning this work allows for the calculation of the predicted informative value of evidence items, for which we calculate the Kullback-Liebler Divergence (KL-D; a measure of entropy reduction; Kullback & Liebler, 1951)[3].

$$KL(E_j) = \sum P(h_i|e_j) * \log\left(\frac{P(h_i|e_j)}{P(h_i)}\right)$$

where E_j is a set of items of evidence $\{E_1, E_{2...}E_j\}$, e_i the set of possible states of the evidence, $\{e_1, e_2, e_i\}$, and h_i is a set of hypotheses, $\{h_1, h_{2...}h_i\}$. In the present case, we compute the information provided by R2 in reference to the hypothesis (H; given we have already observed R1) when a) R2 also relies on E1, vs b) R2 relies on E2, taking a difference measure between these two values.

As such, in asking lay reasoners for their preference for a forthcoming report to be based on shared evidence (i.e. based on an item of evidence already informed by one report) or new evidence, we may observe whether lay reasoning (if in line with normative expectations) predicates an evidence selection preference for more diverse items.

In sum, the present work uses a BN formalism to disentangle the structural vs observation-based forms of shared evidence dependencies. In so doing we are able to not only establish an empirical baseline of when the two forms agree (and thus whether reasoners fit with standard normative expectations), but also examine how reasoners deal with cases of disagreement (where observations appear uncorrelated, but a structural dependence remains), and use structural relations to determine (diversity-based) evidence preferences.

Method

Participants 200 US participants were recruited and participated online through Amazon Mechanical Turk. Three participants were removed for incomplete data, and 1 for not being a native English speaker. Of the 196 remaining participants, 84 identified as female, and the median age was 34 (*SD* = 9.8). All participants gave informed consent, and were paid for their time (*Mdn* = 8.74 minutes, *SD* = 6.63).

Procedure & Design Participants were presented with a scenario in which *a patient, "RN", may have a disease "MTL"* ("H" in Fig. 1). The participant is placed in the role of a diagnostician, attempting to confirm the above diagnosis. They are informed that the patient has had a number of *cell samples* taken (these can be considered $E_{1\text{-}3}$ in Fig. 1), both independently, and of equal diagnosticity. More precisely, that each *cell sample* may contain a *biomarker*, which has a 90% chance of being due to MTL

[1] I.e. P(E1|H) = P(E2|H) = P(E3|H), and P(E1|¬H) = P(E2|¬H) = P(E3|¬H).

[2] I.e. P(R1|E1) = P(R2|E1) = P(R3|E2) = P(R4|E3), and P(R1|¬E1) = P(R2|¬E1) = P(R3|¬E2) = P(R4|¬E3).

[3] Other information measures exist, such as impact (see Nelson, 2005), information gain (Lindley, 1956), and Bayesian diagnosticity (Good, 1950), though empirical work suggests such measures are highly correlated (Nelson, 2005), and are thus considered interchangeable for the present work.

(hit rate), but also a 10% probability of being a false positive.

Participants are then informed that they are unable to examine the *cell samples* themselves, but must rely on *lab technicians* (R1-4 in Fig. 1), who will independently examine the *cell samples* and provide a *report of whether biomarkers are present or absent*. Crucially, all the lab technicians are indicated as equally reliable, in that they have an 80% chance of detecting and reporting a biomarker (irrespective of whether it is due to MTL), when a biomarker is present (hit rate), and a 20% chance of a false positive.

Lastly, participants were informed that prior to receiving any reports from their lab technicians, given the facts of the case so far, they should assume a prior probability of patient RN having MTL of 50% ("Finally, *prior to getting the reports*, you can assume an initial probability of 50% that patient RN has MTL, based on the facts of the diagnostic process so far… Before you start finding out reports, please answer the following question … What is the probability that **patient RN has MTL**?"). This prior probability was then immediately elicited from participants, for use in individual model fitting (see results section below).

Elicitation Stages Participants then received reports from each of 4 lab technicians in turn (resulting in a total of 4 elicitation stages). Following each new report, participants were asked to provide a new probability estimate of patient RN having MTL – given *everything they now know* (i.e. background + gradually accumulating reports). These probability estimates were the main dependent variable.

Each report statement took the form "*Based on their assessment of* ***cell sample [1/2/3]****, lab tech [1/2/3/4] reports that the* ***biomarker is [present/absent]****.*"

Crucially, there were two independent, between-subject variables employed, making a 2x2 design. The first of these was the *evidence used by the second lab technician* ("R2Evidence"). Whilst the first lab technician always used cell sample 1, the third cell sample 2, and the fourth cell sample 3, the second lab technician used cell sample 1 in one condition (R2E1), and cell sample 2 in the other (R2E2). This allowed for a) the between-subject comparison of 2 reporters using independent (R2E2) vs shared evidence (R2E1), and b) allowed for the disentanglement of *structure* (i.e. dependency relations) from *order of observations* (i.e. is over/under updating due to the second report relying on shared evidence, or simply because it is the *second* report).

The second between subject factor was the order of positive (biomarker present) and negative (biomarker absent) reports ("RepOrder"). More precisely, either the first lab technicians 1 and 2 gave positive reports (and 3 & 4 gave negative reports; "PosFirst"), or the reverse ("NegFirst"). This general structuring, when taken in conjunction with the R2Evidence factor, allowed for the assessment of the influence of shared evidence when reporters agree about the same evidence (R2E1) or disagree (R2E2 – as lab technicians 2 & 3 will always disagree, yet will share cell sample 2). Additionally, this allows for the further disentanglement of observation *type* from shared evidence (structural) influences. For instance, in R2E1 conditions, the reports from shared evidence (lab technicians 1 & 2) will half the time be positive, and the reports from independent evidence (lab technicians 3 & 4) will half the time be negative, and vice versa. Thus, one may discern the influence of (dis)confirming observations vs structural differences.

Dependent variables Along with the probability estimates elicited at each elicitation stage (0-100% slider, no default)[4], one further qualitative question was asked after the first lab technician provided a report (i.e. elicitation stage 1):

"Given the choice, would you rather *Lab Tech 2* **also independently investigated cell sample 1 for a biomarker, or investigated a different cell sample (cell sample 2)**?" ["Same cell sample (cell sample 1)" / "Different cell sample (cell sample 2)" / "There is no difference."] Forced choice, randomized order of presentation.

The purpose of this question was to assess participant preferences for diversity (independence in this case) in their observations.

Taken together, the probability estimates and evidence preference judgment allow for the assessment of the impact of shared evidence, both in terms of predicted support, and consequent reasoning (and belief-updating), whilst taking into account the influence of observation types and orders.

To concretize the research questions into hypotheses, we predict:

H1. Shared Evidence Structure - Shared evidence will result in estimates of reduced impact of affected reports, in comparison to reports from distinct items of evidence, *ceterus paribus*. Tested via the between subject comparison of the impact of *lab technician 2* at the second elicitation stage when *lab technician 2* does/does not share *cell sample 1* with *lab technician 1* (i.e. R2E1 vs R2E2 conditions).

H2. Contradictions and Dependence – Reasoners will find the integration of two contradicting reporters using the *same* evidence (i.e. *within* a shared evidence item) more difficult than when contradictions are based on separate evidence items. Tested via the comparison to normative expectation at third elicitation stage (when *lab technician 1, lab technician 2,* and *lab technician 3* have reported) in R2E2 condition in comparison to R2E1 condition, where contradiction cuts across (rather than within) evidence items. Such a prediction is informed by previous work that has found lay reasoners struggle with inferences from contradicting reports across a dependency (Pilditch, Hahn, & Lagnado, 2018).

H3. Diversity Preference - Participants (correctly) prefer more diverse evidence (prefer *lab technician 2* to use evidence *cell sample 2*.

i. An additional question of interest is whether diversity preferences will be lower when *lab*

[4] Open text reasoning responses were also collected at the end of each elicitation stage, but for the sake of brevity are not reported here.

technician 1 provides negative evidence (NegFirst condition), than when *lab technician 1* provides positive evidence (PosFirst)?

Fig. 2 below shows the different structural and report order comparisons for the 2x2 design (each cell is a between subject condition), with T1 to T4 within each cell as the within-subject order of evidence. Thus, H1 is investigated by comparing T2 in the top row (when R2 is reliant on the same evidence as R1) with T2 on the bottom row (when R2 is using different evidence). We can then assess H2 by comparing T2 to T3 in the top row (contradiction based on separate items) to the bottom row (contradiction based on shared evidence. H3 is assessed having seen the report at T1 (and is asked prospectively about T2), and H3i. is based on the comparison of responses to the H3 question in left versus right columns of Fig. 2.

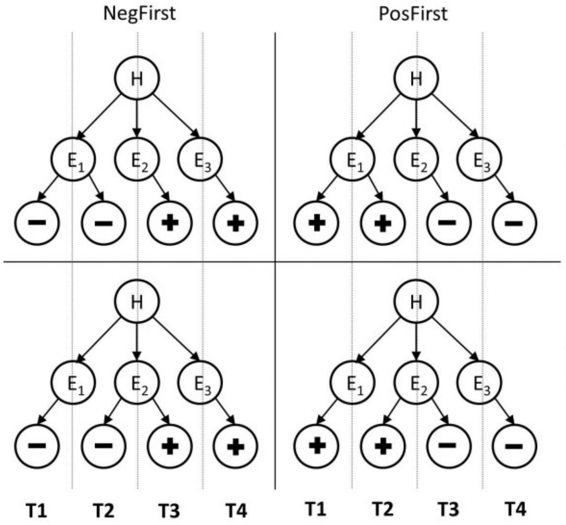

Figure 2. Underlying networks, split by R2Evidence (rows) and RepOrder (columns) conditions. T1 to T4 reflect sequence of reports (within-subjects).

Results

Bayesian statistics were employed throughout[5] using the JASP statistical software (JASP Team, 2018). Using the gRain package in R (Højsgaard, 2012), the elicited priors from each participant were used to individually fit BNs for each participant. Remaining parameters were as specified in the background information presented to participants. The posterior probabilities at each elicitation stage generated from each BIBN model (representing each participant) were used in subsequent comparison analyses.

Probability Estimates

The hypothesis-directed analyses used to unpack a) the influence of when shared evidence is introduced (*H1*), and b) the influence of contradiction within/outside a dependency (*H2*), first employed an RM-ANOVA on participant estimates alone (including between subject factors), so as to determine participant behavior, followed by a further analysis that compared these estimates to BIBN predictions, to determine the "correctness" of this behaviour.

H1. Firstly, to assess *H1*, an RM-ANOVA on participant estimates from T1 to T2, found participants were insensitive to R2Evidence condition overall, $BF_{Inclusion}$ = 0.102, or in interaction with elicitation stage, $BF_{Inclusion}$ = 0.105. This was despite participants updating in light of new evidence *in general*, $BF_{Inclusion}$ > 10000, and whether that evidence was positive or negative, $BF_{Inclusion}$ > 10000. This was further evidenced by the interaction of elicitation stage and RepOrder (participants decreased estimates as negative reports came in, and increased as positive reports came in), $BF_{Inclusion}$ > 10000. Consequently, the model of participant estimates without R2Evidence yielded the strongest fit, BF_M = 63.2, and was decisive overall, BF_{10} > 10000.

Consequently, by subsequent inclusion of the BIBN predictions for each participant (the Observed vs Predicted factor), this insensitivity to shared evidence (i.e. the influence of R2Evidence, was shown to be insufficient relative to (fitted) normative expectation. This was evidenced by a main effect of Observed vs Predicted, $BF_{Inclusion}$ = 5479.38, and critically, strong evidence for the interaction of R2Evidience and Observed vs Predicted (BIBN predictions changed with R2Evidence, whilst participant estimates do not), $BF_{Inclusion}$ = 11.72.

In sum, these analyses revealed participants were insensitive to impact of shared evidence structures, as compared to their fitted Bayesian predictions (*H1*).

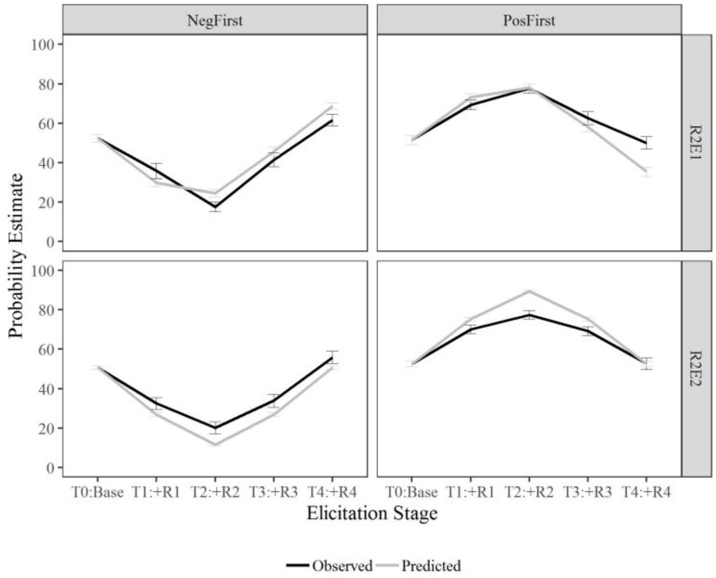

Figure 3. Probability estimates across elicitation stages, split by condition. Error bars reflect standard error.

H2. Secondly, to assess *H2*, the same analytical protocol was used on elicitation stages T2 to T3. This corroborated H1 findings, in that the insensitivity to the influence of

[5] For all analyses, an uninformed prior was used. Wherever possible, sample sizes for a given analysis (*N*), and Bayesian Credibility Intervals (95% CI) are indicated.

R2Evidence (this time via the presence of shared evidence in concert with contradicting reports either outside, R2E1, or within, R2E2, the same shared evidence) persisted both overall, $BF_{Inclusion}$ = 0.315, and in interaction with elicitation stage, $BF_{Inclusion}$ = 0.321. However, once again participants were sensitive to the introduction of new evidence in general, $BF_{Inclusion}$ > 10000, its valence, $BF_{Inclusion}$ > 10000, and the interaction of these factors (newly introduced positive reports lead to increased estimates, whilst newly introduced negative reports lead to decreased estimates), $BF_{Inclusion}$ > 10000. As with the *H1* analysis, the model of participant estimates without R2Evidence yielded the strongest fit, BF_M = 20.425, and was decisive overall, BF_{10} > 10000.

To again determine whether this insensitivity was erroneous, BIBN predictions for each participant were included as another within subject factor (Observed vs Predicted). Again, participant estimates were shown to not only be generally insufficient in comparison to BIBN predictions, $BF_{Inclusion}$ > 10000, but that this insensitivity extended to shared evidence (R2Evidence x Observed vs Predicted; BIBN estimates change with condition, participant estimates do not), $BF_{Inclusion}$ > 10000.

In conclusion, the above analyses corroborate the insensitivity findings of *H1*, extending them to the issue of contradiction (of reports) being based on the same or different evidence items (*H2*).

Taken together, *H1* and *H2* findings suggest participants were insensitive to the impact of shared evidence, both when reporters are corroborating with, and contradicting each other.

Evidence Preference

The BIBN models for each participant, having taken into account the elicited prior for the hypothesis, generated the expected information gained in KL-D, having observed the positive/negative report from the first lab technician, for two models; one in which the second lab technician used the same evidence as the first (E1), and one where the second lab technician used different evidence (E2). The difference in expected information gain between these two models was used to generate a normative preference (based on maximum expected information) for the second lab technician using E1, E2, or them being equivalent ("NoPref").

To assess the observed evidence preferences, a Bayesian binomial test was conducted on observed preferences (dark grey bars of Fig. 4), comparing them to chance responding (0.33). Preferences for the second lab technician to use the *same* evidence as the first lab technician (E1) were found to be at chance level (0.36, 95% CI: [0.293, 0.426]; *N* = 196), BF_{10} = 0.118, whilst diversity preferences (second lab technician to use E2) were found to occur decisively above chance (0.51, 95% CI: [0.441, 0.579]; *N* = 196), BF_{10} > 10000, lending some support to the diversity preference predicted (*H3*). The frequency of participants opting for "no preference" was decisively below that expected by chance, (0.13, 95% CI: [0.092, 0.187]; *N* = 196), BF_{10} > 10000. A Bayesian contingency table revealed these frequencies to not be influenced by whether the first lab technician had made a positive (right-hand facet of Fig. 4) or negative (left-hand facet of Fig. 4) report (*N* = 196), BF_{10} = 0.045, speaking against hypothesis *H3i*.

Crucially, participant preferences for the second lab technician to use the *same* evidence as the first lab technician are substantially higher than that predicted by BIBN models (i.e. 0; see light grey bars of Fig. 4). This is corroborated by the decisive deviation in frequencies between observed and predicted preferences (*N* = 392), BF_{10} > 10000. Put another way, and contrary to predictions of *H3*, approximately 1/3rd of participants retain an explicit preference for the information-poorer reports that "confirm" (i.e. are based on evidence that has already formed the basis of an observed report), rather than a diversity preference or lack of preference.

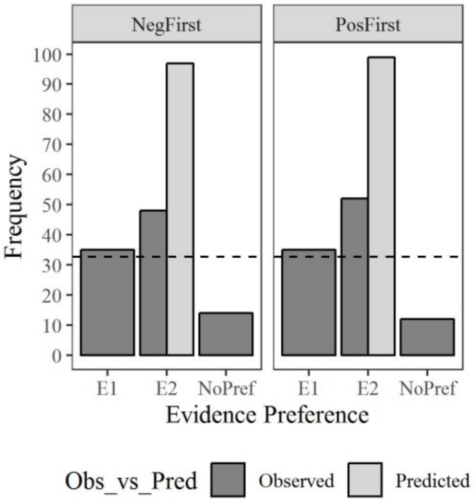

Figure 4. Evidence Preferences, split by condition. Dashed line represents chance level (33%).

Conclusions

When reasoning under uncertainty, an important consideration is the impact of dependencies among evidence items. More precisely, seemingly separate reports, which in fact stem from the same source (or evidence basis), carry redundant information, relative to truly separate reports (based on distinct information). To mistake the former for the latter can lead to overweighting support for a given hypothesis, to deleterious consequences (Dror et al., 2006; Koller & Friedman, 2009).

Here, we show that lay reasoners seem rather insensitive to the impact of this form of dependency and consider the two cases equivalent when estimating degrees of support for a hypothesis. At the same time, our findings corroborate prior research in terms of both a) the consistent under-weighting of introduced evidence (see e.g. Faigman &

Baglioni, 1988; Nance & Morris, 2005), and b) more substantial deviations when having to deal with contradictory reports (Pilditch et al., 2018).

Finally, we present a second novel finding in lay reasoners preferences for further reports based on shared (i.e. previously informed upon) evidence (a "confirmatory" preference) or separate (unseen) evidence (a "diversity" preference). Though the majority of participants conform to a diversity preference in line with maximising expected information, approximately 1/3rd of lay reasoners have a confirmatory preference. While failures to appreciate diversity have been reported before (e.g., Soll, 1999), there are clear preferences for diversity in other inferential contexts (e.g., Rips, 1979; Osherson et al., 1990), even in children (Heit & Hahn, 2001). Hence further work will be required to pinpoint exactly for when, where and why diversity is appreciated and when it is not. It is worth note in this context that where the reliability of the reporting sources is not exactly known (unlike the lab technicians in the present study), diverse evidence is arguably not always normatively superior (see Bovens & Hartmann, 2003). Whether lay reasoners have any understanding of the different circumstances where a diversity advantage does and does not obtain remains to be seen.

References

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press on Demand.

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic science international*, 156(1), 74-78.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahneman, D., Slovic, P., & Tversky, A. (eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press. pp. 249--267.

Fenton, N., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. Crc Press.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*(1), 61-102.

Faigman, D. L., & Baglioni Jr, A. J. (1988). Bayes' theorem in the trial process: Instructing jurors on the value of statistical evidence. *Law and Human Behavior*, *12*(1), 1.

Good, I. J. (1950). *Probability and the weighing of evidence*. New York: Griffin.

Harris, A.J.L., & Hahn, U. (2009). Bayesian Rationality in Evaluating Multiple Testimonies: Incorporating the Role of Coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1366–1372.

Heit, E. H. U. (2001). Diversity-Based Reasoning in Children. *Cognitive Psychology*, *43*, 243–273.

Heuer, R. J. (1999). *Psychology of intelligence analysis*. Washington DC: Center for Study of Intelligence.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1-26.

Hogarth, R. M. (1989). On combining diagnostic ''forecasts'': Thoughts and some evidence. *International Journal of Forecasting,* 5, 593–597.

JASP Team (2018). JASP (Version 0.9)[Computer software].

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, *7*(4), 317-337.

Kullback, S., & Liebler, R. A. (1951). Information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

Lagnado, D. A. (2011). Thinking about evidence. In *Proceedings of the British Academy* (Vol. 171, pp. 183-223).

Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27, 986–1005.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2018). Partial source dependence and reliability revision: the impact of shared backgrounds. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 722-727). Austin, TX: Cognitive Science Society.

Nance, D. A., & Morris, S. B. (2005). Juror understanding of DNA evidence: An empirical assessment of presentation formats for trace evidence with a relatively small random-match probability. *The Journal of Legal Studies*, *34*(2), 395-444.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112(4), 979–999.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2009). *Causality. Models, reasoning, and inference*. Second edition. New York: Cambridge University Press.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*(2), 242.

Phillips, K., Hahn, U., & Pilditch, T. D. (2018). Evaluating testimony from multiple witnesses: single cue satisficing or integration? In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2244-2249). Austin, TX: Cognitive Science Society.

Pilditch, T.D., Hahn, U., & Lagnado, D. (2018). Integrating dependent evidence: naïve reasoning in the face of complexity. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference* *of the Cognitive Science Society* (pp. 884-889). Austin, TX: Cognitive Science Society.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Memory and Language*, *14*(6), 665.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Northwestern University Press.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 105-151.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*(2), 317-346.