Integrating dependent evidence: naïve reasoning in the face of complexity

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

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK
[2]Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK

Abstract

When reasoning about evidence under conditions of uncertainty, one important consideration for accurate updating is the presence (and influence) of dependencies. For instance, if considering whether a patient has a disease, the value of two doctors' diagnoses indicating the presence of the disease may carry more value if such diagnoses were conducted independently, rather than if, all else being equal, one doctor has seen the other's diagnosis before making their own. In the present paper, we demonstrate that lay reasoners prefer to avoid dependencies when considering evidential support. However, we additionally illustrate two cases in which dependencies may carry evidential *advantage:* namely, when information is partial or contradictory. Lay reasoners erroneously remain averse to dependencies even in such cases, reflecting the difficulties inherent to considerations of dependence.

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

Introduction

The accurate integration of information is essential to everyday life, be it in order to reason effectively, to come to conclusions, or to make decisions. The accuracy of this integration has been of particular concern in the domains of intelligence analysis (Heuer, 1999), law (Fenton & Neil, 2012, Fenton, Neil & Lagnado, 2013, Hahn & Oaksford, 2007; Harris & Hahn, 2009; Lagnado, 2011; Pennington & Hastie, 1986; Schum, 1994), and medicine (Eddy, 1982), where efforts have been made not only to understand how people *do* integrate, but also how they *should* do so. Questions remain, from both empirical and normative standpoints, regarding the central issue of how to deal with dependencies *between pieces of evidence*.

In modelling the integration of evidence, it is sometimes legitimate to assume conditional independence (see e.g., Bovens & Hartmann, 2003; Hahn, Harris & Corner, 2009; Fenton & Neil, 2012), such that knowledge of one piece of evidence does not affect the impact of another (Pearl, 1988). Such an assumption eases computation (Schum, 1994; Pearl, 1988), but if misapplied can lead to dangerous over-weighting of evidential value (e.g., naïve Bayes in medicine; Koller & Friedman, 2009; Kononenko, 1993).

To illustrate, Bayesian networks (BNs) provide a graphical, computational framework for reasoning under uncertainty, using Bayes theorem and conditional likelihoods to make optimal inferences (Pearl, 1988; 2009). For the purpose of this paper, we are interested in the simple case of a *unidirectional, direct relation* between two reporters. In other words, a dependency characterised by one reporter (S_A) "receiving" information from another reporter (S_B) *before* providing their own report about the *same* hypothesis (H; see Fig. 1). Such a structure is relevant to many areas of enquiry, including medicine (e.g., a doctor may make a diagnosis having already seen the diagnosis of a second doctor), and the legal / forensic domain (e.g., a lab technician runs a finger print analysis already knowing exonerating information about the suspect; see Dror, Charlton, & Peron, 2006).

If the two sources corroborate one another, and are conditionally independent (no dashed line), then their reports carry more weight than if the two sources have colluded, shared information (dashed line), or possess some other form of dependence (all else being equal). This is because in the dependent case there are alternative explanatory paths for a source's report – other than the true state of the world: for instance, a doctor (S_A) diagnoses a disease (H) based primarily on the diagnosis of another doctor (S_B), rather than their own independent assessment of the symptoms. Conversely, sources independently providing corroborative reports exert a direct, multiplicative impact on H (in accordance with standard Bayesian updating; Pearl, 1988). Accordingly, to mistake dependent structures for independent (or vice versa) can lead to systematic over (or under) weighting of evidential value.

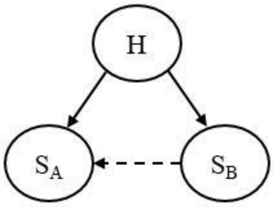

Figure 1. Graphical representation of a hypothesis (H) with two sources of evidence (S_A, S_B) informing upon it.

Judgments of dependence, and the case for dependency advantages

When multiple reports corroborate one another, then (all else being equal) independent reports will convey more support than those sharing dependencies. This notion fits with conceptualisations of dependence-as-redundancy in expert forecasting (Hogarth, 1989; Soll, 1999). However, there are instances in which independence is *not* preferable. In this paper, we introduce two such instances, and experimentally assess lay reasoners intuitions about the impact of dependence in such cases.

To illustrate, consider the two columns of Fig. 2: the networks in the two cases are both of two sources (A and B) reporting on a hypothesis of interest (H). The cases differ in the additional arrow indicating a dependency between the two sources (left-hand column). At t_0, neither source has provided a report. Imagine now receiving only the report of Source A (row t_1; indicated by the 100% in row "H" for the Source A cell, reflecting a report that H is true). In this state of 'partial information', the dependent case (left) can lend more support to the hypothesis (H) than the independent counterpart (right) – as seen by the higher resulting probability of H (84% vs 80%).[1] Imagine now receiving the second report, which *contradicts* the first (Source B reports that H is false; t_2 of Fig. 2). Here dependencies can also lead to greater support for H than the independent case (60% vs 50%). Of course, if reports corroborate at t_2 (not shown), independence always yields more support than dependence (94% vs 90% in this example). In this paper, we probe lay reasoners understanding of such situations.

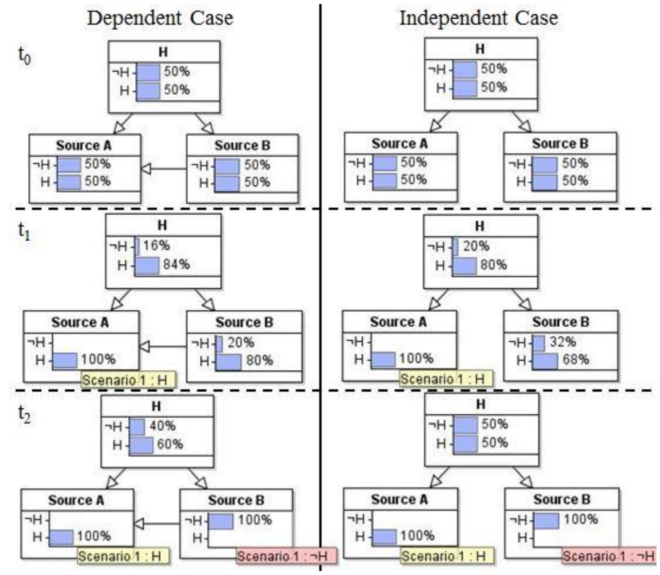

Figure 2. Common-cause structures, with independent and dependent cases, across no information (t_0), partial (t_1) and contradicting information (t_2) states.

Present research We ask how lay reasoners understand and use judgments of the evidential impact of dependencies. For instance, do people assume that dependencies imply lower informative value, all else being equal? Does such an assumption extend to cases in which dependence is in fact preferable?

The advantage a dependency provides relies not only on the pattern of evidence (e.g. contradictory vs. corroborative), but also on the manner in which error rates of an independent source are modified, given access to a secondary source (i.e. the change in reliability of S_A (Fig. 1) when the (dashed) line of dependence from S_B is introduced). For instance, to what degree are error rates mitigated when a secondary source provides accurate information (or exacerbated by false information)? We elicit these modifications (conditional probabilities) from participants, which allow us to determine the coherence of reasoners in judging the impact of dependencies by their own lights. This approach also affords an opportunity to explore lay intuitions regarding the impact of secondary sources on the reliability of a recipient source.

Method

Participants 200 US participants were recruited and participated online through Amazon Mechanical Turk. Participants were native English speakers[2], with a median age of 33 (*SD* = 11.06), and 107 participants identified as female. All participants gave informed consent, and were paid for their time.

Procedure & Design Participants were presented with background information describing a plane crash that may have been caused by sabotage (P(H) = .5), for which there were two experts, Bailey (Source A in Fig. 2) and Campbell (Source B in Fig. 2), who independently assessed the crash site (so as to provide reports about the hypothesis of sabotage). Both experts were indicated to be reasonably reliable (with error rates – both false positive and false negative – of 20%[3]) when independent.

Participants then provided conditional probabilities for Bailey's adjusted reliability when given (in)correct information from Campbell (used in model comparison; see below), followed by three elicitation stages in which participants compared the support for the hypothesis provided in dependent versus independent scenarios as gradually more information was presented (see questions and scenarios below). These stages follow those illustrated in Fig. 1: t_0(Baseline) – no reports; t1(First report) – Bailey gives a positive (sabotage) report; t_2(Second report) – Campbell gives either a corroborating (sabotage) or contradicting (no sabotage) report. Whether participants saw a corroborating or contradicting t_2 report was manipulated between-subjects. Conditional probability questions, scenarios, and elicitation stage questions are shown below.

Conditional probability questions To inform model comparisons, participants were first asked to provide probability estimates (0-100%) for the following two conditions:

1. *"If Bailey, before making her report, has seen Campbell's completed report - when that report is in fact CORRECT - what do you estimate is the probability of Bailey making a mistake now?"*

[1] This depends on the complex relationship between independent and dependent source error rates, as well as P(H). In the case shown in Fig. 2, P(H) = .5, and independent sources have error rates of .2 (i.e. false positives = false negatives). The issue of how to alter error rates for a dependent source remains an open question from a normative standpoint, but for the present example, error rates are assumed to halve if provided with an accurate secondary report, and double if provided with a misleading report. Thus, $P(\neg Rep_A|Rep_B,H) = .1$, whilst $P(\neg Rep_A|\neg Rep_B,H) = .4$.

[2] 202 participants were originally recruited, but 2 were removed for not being based in the US nor being a native English speaker.

[3] I.e. $P(Rep_B|\neg H) = P(\neg Rep_B|H) = P(Rep_C|\neg H) = P(\neg Rep_C|H)= .2$

2. *"If Bailey, before making her report, has seen Campbell's completed report - when that report is in fact INCORRECT - what do you estimate is the probability of Bailey making a mistake now?"*

The two questions thus elicit from participants an estimate of the change in error rates for a recipient source (Bailey) given exposure to a secondary "sending" source (Campbell) when the latter is in fact 1) correct ($P(\neg Rep_B|Rep_C,H)$), or 2) incorrect ($P(\neg Rep_B|\neg Rep_C,H)$).

Scenarios For the three elicitation stages, participants were given two scenarios to consider (with the background re-presented):

"Scenario 1: You learn that Bailey, prior to completing her report, was accidentally given access to Campbell's completed report. As such, Bailey's report may be influenced by what Campbell has reported."

"Scenario 2: You learn that Bailey completed her report without ever seeing Campbell's completed report. As such, Bailey's report is not influenced by what Campbell has reported."

Scenario 1 is thus a "dependent" scenario, and Scenario 2 an "independent" scenario.

Elicitation Stage questions The three questions pertaining to the scenarios across the three elicitation stages were:

Qualitative Judgment (forced choice): *"Based on what you know at this point, which scenario (if either) provides more support for the plane having been sabotaged?"* ["They are the same." / "Scenario 1" / "Scenario 2". Randomized presentation order.]

Confidence in Judgment: *"How confident are you that your response is correct?"* [Slider, 0 – 100%.]

Probability Estimates: *"What is your current probability estimate of sabotage in each scenario, given what you know so far?"* [Sliders from 0-100%.]

Open text reasoning was also elicited after each stage, but for the sake of brevity is not reported here.

Results

Bayesian statistics were employed throughout[4] using the JASP statistical software (JASP Team, 2017). For the sake of brevity, analyses are not reported exhaustively here.

Elicited conditional probabilities As illustrated in Fig. 3, participants appear to generally reduce a source's error rates when that source has been provided with correct information from a secondary source (median (green dashed line in Fig. 3) = 15%), and increased error rates when that secondary source is incorrect (median (red dashed line in Fig. 3) = 36%).

A Bayesian T-Test was conducted on each conditional probability to assess deviation from the starting value of 20%. Estimates of the impact of an "incorrect" secondary source showed decisive evidence for a difference (N = 200; M = 38.05, 95% CI: [35.19, 40.9]), BF_{10} = 3.707 * 10^{23}. However, the impact of a "correct" secondary source on error rates showed strong evidence for a null difference (N = 200; M = 20.73, 95% CI: [18.13, 23.34]), BF_{10} = 0.092.

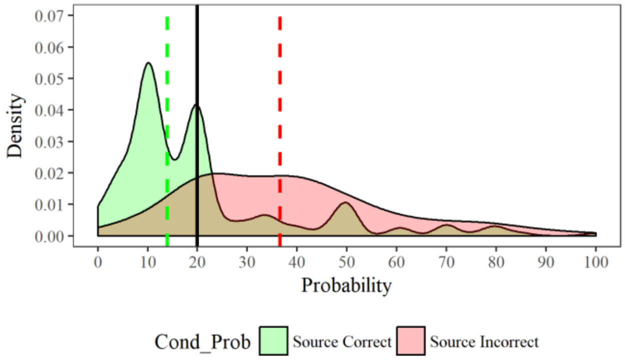

Figure 3. Elicited conditional probabilities of the expected error rate for a dependent source with a standard (independent) error rate of 20%, when provided with correct (green) or incorrect (red) second-hand information.

This asymmetry (i.e. greater revision upwards than downwards) is expected given the "generally reliable" starting (independent) error rates (because, given the bounded scale, an initial error rate of 20% can be increased more than it can be decreased). It is additionally worth noting that the large variance exhibited in elicited conditional probabilities may be reflective of participant interpretations of the impact of *background information* (Schum 1994) - inherent to considerations of dependence.

Behaviorally Informed Bayes Net (BIBN) models Using the gRain package in R (Højsgaard, 2012), the elicited conditional probabilities from each participant were used to outfit the error rates for Bailey (as a recipient / dependent source) in a dependent-scenario BN, creating individually fitted BNs for each participant. Both the priors and *independent* source / scenario probabilities were as specified in the background information presented to participants. The posterior probabilities and qualitative judgments for each scenario (at each elicitation stage) generated from each BIBN model (representing each participant) were used in subsequent comparison analyses.

Qualitative judgments

To analyze qualitative judgments, a series of Bayesian contingency tables (to assess factors) and binomial tests (for chance level comparisons) were used. These first assessed participant judgments across scenario types (dependent vs independent; left vs right columns of Fig. 2), elicitation stages (Baseline, First Report, Second Report; t_0, t_1, and t_2 in Fig. 2), and conditions (contradicting vs corroborating

[4] According to Jeffreys (1961), Bayes Factors (BF_{10}: likelihood ratio of data given hypothesis, over data given null), may be interpreted as: 1 – 3 = anecdotal support; 3-10 = substantial; 10-30 = strong; 30-100 = very strong; >100 = decisive. Conversely, Bayes Factors < .33 can be considered strong support for the *null* (Dienes, 2014). For all analyses, an objective (uninformed) prior was used, unless otherwise specified. Wherever possible, sample sizes for a given analysis (N), and Bayesian Credibility Intervals (95% CI) are indicated.

second reports), before comparing them with participant BIBN model predictions.

Participant data (dark grey bars, Fig. 4). A judgment (3) x elicitation stage (3) contingency table (*N* = 600), found strong evidence for an effect of elicitation stage on participant judgments, BF_{10} = 5.724, indicating sensitivity to elicitation stage. An additional judgment (3) x condition (2) contingency table on the second report judgments only (where the condition – corroborating or contradicting – occurred) found very strong evidence for the effect of condition (*N* = 200), BF_{10} = 31.97, reflecting sensitivity to whether the second report contradicted or corroborated the first.

BIBN model (light grey bars, Fig. 4) comparison. To compare participant judgments to BIBN model predictions, a series of Bayesian contingency tables using a "data type" (Participant vs. BIBN prediction) factor were employed across the three elicitation stages (separating contradicting and corroborating conditions in the second report stage).

Decisive evidence was found for the deviation of participant judgments from those predicted by BIBN models across baseline (*N* = 400), $BF_{10} = 3.588 * 10^{25}$, first report (*N* = 400), $BF_{10} = 9.001 * 10^{6}$, second report-corroborating (*N* = 200), BF_{10} = 81188.756, and second report-contradicting (*N* = 200), $BF_{10} = 1.265 * 10^{6}$, stages.

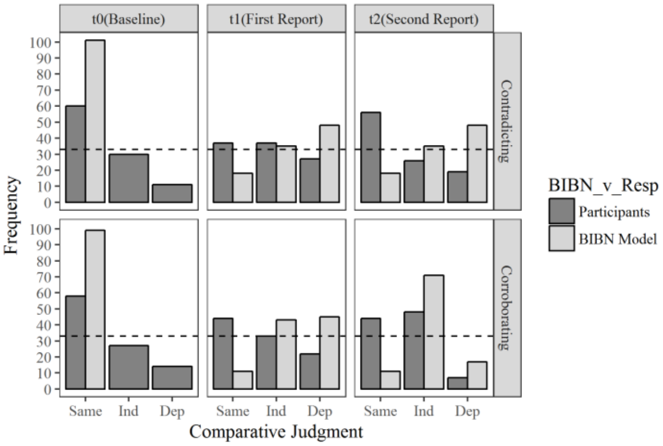

Figure 4. Qualitative comparison judgments, split by elicitation stage and condition. Dashed line represents chance level (33%).

To determine the degree of internal coherence (individual level BIBN model prediction vs actual judgment), a variable was created for each participant judgment. If a judgment corresponded to that predicted by the BIBN model for that participant, then the judgment was considered correct (1), or else incorrect (0). This "coherence" variable was then used to determine if participants made correct judgments more often than expected by chance (0.33) across elicitation stages and conditions, using binomial tests.

Correct responding occurred above chance level at baseline (0.59, 95% CI: [0.521, 0.656]; *N* = 200), $BF_{10} = 1.678 * 10^{11}$, but strong evidence for the null (no difference from chance) was found for responses at first report (0.34, 95% CI: [0.278, 0.408]; *N* = 200), BF_{10} = 0.088.

Breaking down the second report by condition, substantial evidence was found for correct responding above chance level in the corroborating condition (0.455, 95% CI: [0.36, 0.553]; *N* = 100), BF_{10} = 3.388, whilst strong evidence for the null was found in the contradicting condition (0.356, 95% CI: [0.27, 0.454]; *N* = 100), BF_{10} = 0.139.

Taking these results together, participant judgments were generally inconsistent with BIBN predictions. Correct response rates were greater than chance in baseline and corroborating cases, but no better than chance in cases of partial (first report) and contradicting (second report) information – situations in which (by participants own lights) dependencies may be advantageous (see light-grey bars in center and top-right panels of Fig. 4).

Confidence in qualitative judgments. Confidence was generally high across all judgments (*M* = 67.49, *SD* = 23.84). Although it was not affected by judgment category or elicitation stage[5], confidence was decisively higher when second reports (*N* = 200) corroborated (*M* = 74.4, *SD* = 23.61) rather than contradicted (*M* = 60.88, *SD* = 25.26), BF_{10} = 163.385. This finding fits with the higher erroneous responding found in qualitative judgments when evidence is contradictory.

Probability estimates

We next turn to participant probability estimates. The purpose of this analysis was to determine a) the manner in which participants are updating quantitatively in light of new information (and whether this differed across independent and dependent scenarios) and b) how this updating compares to BIBN model performance – for instance, highlighting insufficient support attributed to dependent scenarios.

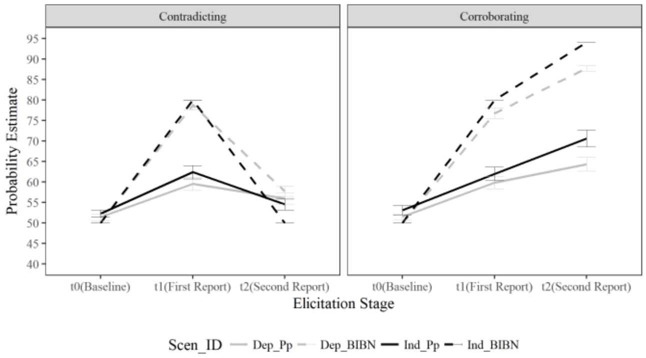

Figure 5. Probability estimates across elicitation stages, split by condition. Error bars reflect standard error.

[5] A Bayesian ANOVA assessing the impact of judgment and elicitation stage (*N* = 600) on confidence found no evidence for the effect of judgment, $BF_{Inclusion}$ = .489, and strong evidence for the null for the effect of elicitation stage, $BF_{Inclusion}$ = .014. The interaction term also showed strong evidence for the null, $BF_{Inclusion}$ = .001.

Participant data (solid lines, Fig. 5). To determine the manner of participant responding, a series of Bayesian repeated-measures ANOVA were used. Although not reported exhaustively here for space reasons, accompanying footnotes provide necessary supplementary statistics.

Overall, independent scenarios (black solid lines, Fig. 5) were considered to provide more support than dependent equivalents (grey solid lines, Fig. 5), $BF_{Inclusion}$ = 7.269[6], substantiating qualitative judgment findings. Notably, in the (second report) corroborative case, participants assigned greater support to the independent scenario (relative to dependent), BF_{10} = 223.233, whilst in the contradicting case there was a null difference, BF_{10} = 0.162[7]. This again fits the qualitative judgment data, wherein the corroborative case (which is typified by independent scenario advantage) is easier for participants to determine (confidence was also higher).

Finally, there was a main effect of elicitation stage (linearly increasing estimates across stages), $BF_{Inclusion} > 150$, indicating participants were generally sensitive to incoming information, including the impact of contradictory vs corroborating information, $BF_{Inclusion} > 150$[8]. To address the question of how *sufficient* this updating is, participant data was compared to BIBN model predictions.

BIBN model (dashed lines, Fig. 5) comparison Although Fig. 5 shows the aggregates for BIBN model predictions, to appropriately explore model fit on the individual level, a Bayesian repeated measures ANOVA was conducted on probability estimates with the additional inclusion of data type (Participant vs. BIBN prediction) as a within-subject factor (N = 2400).

Decisive evidence was found for the main effect of data type, $BF_{Inclusion} > 150$, indicating probability estimates were significantly higher in BIBN model predictions. Decisive evidence was also found for the interaction of data type with elicitation stage (increasing deviation over stages), $BF_{Inclusion} > 150$, condition (greater deviation in corroborating), $BF_{Inclusion} > 150$, and their 3-way interaction (the greater deviation in corroborating occurs in second report state), $BF_{Inclusion} > 150$[9]. Taken together, this demonstrates that participant updating is insufficient relative to their BIBN model predictions (i.e. evidence is generally undervalued). Further, the interaction with condition (as with participant data alone) motivates the restricted analysis of second report estimates only, split by condition (now also including data type).

In the corroborating condition, decisive evidence was found for the main effect of scenario type, $BF_{Inclusion}$ = 898.016, and data type, $BF_{Inclusion} > 150$, but not their interaction, $BF_{Inclusion}$ = 0.899[10]. This indicates that although participants generally undervalue the impact of a second, corroborative report, their assessment of the greater support provided in independent (relative to dependent) scenarios is broadly in line with model predictions.

Conversely, in the contradicting condition, there was no evidence for the main effect of data type, $BF_{Inclusion}$ = 1.183, and strong evidence for the null for both the main effect of scenario type, $BF_{Inclusion}$ = 0.08, and their interaction, $BF_{Inclusion}$ = 0.039[11]. Although this appears to indicate a reasonable fit between participant data and model predictions for contradicting evidence cases, such a null difference may in fact be attributable to under-valuing evidence *twice*, rather than accurate updating.[12]

Conclusions

The issue of dependence in evidential reasoning, both from empirical and normative standpoints, is critical across many domains, including medical, legal, and intelligence analysis. Failures to consider dependencies can be dangerous, as they can lead to the systematic over-valuing of evidence (Dror et al., 2006; Koller & Friedman, 2009). Our results indicate that lay reasoners are indeed aware of the lower evidential value that dependencies *can* incur. For example, lay reasoners preference for independence is applied appropriately in cases of corroborative reports.

In cases of partial or contradicting information, however, where the implications of dependence are more complex, lay reasoners begin to struggle. More precisely, when by the lights of their own elicited conditional probabilities, dependencies should in fact carry an informational *advantage* (all else being equal) participants still prefer to avoid them. This aversion to dependencies in the qualitative data is corroborated by participant probability estimates.

[6] A Bayesian, repeated-measures ANOVA including all relevant factors (within: scenario type, elicitation stage; between: condition) was run in a hierarchical model (N = 1200). The model that included main effects for elicitation stage, scenario type, and condition, as well as the interactions of elicitation stage x condition and scenario type x condition, enjoyed the strongest support, BF_M = 56.245, with decisive evidence overall, $BF_{10} = 2.619 * 10^{44}$. This model is hereafter referred to as $Model_P$.

[7] Motivated by the significant scenario type x condition interaction in $Model_P$ ($BF_{Inclusion}$ = 8.078), two separate Bayesian ANOVAs were conducted on estimates in the second report elicitation state, split by condition, testing the effect of scenario type in each case (N = 200 in each case).

[8] Decisive evidence was found for the elicitation stage x condition interaction in $Model_P$ (along with a decisive main effect of condition, $BF_{Inclusion} > 150$).

[9] Data type did not interact with scenario type (both in isolation, $BF_{Inclusion}$ = 0.041, or in conjunction with other factors). Accordingly, the model combining $Model_P$ with data type and the significant interactions above yielded the comparatively strongest fit, BF_M = 914.889, with decisive evidence overall, $BF_{10} = 1.485 * 10^{356}$.

[10] Thus, the model consisting of the two main effects only provided the comparatively strongest fit, BF_M = 17.705, with decisive evidence overall, $BF_{10} = 3.539 * 10^{41}$.

[11] Thus, although the model consisting solely of data type was the comparatively strongest fit, BF_M = 5.259, there was no evidence for this model overall, BF_{10} = 1.751.

[12] A Bayesian repeated measures ANOVA of first to second report estimates in the contradicting condition reveals an interaction of data type with elicitation stage from first report to second report, $BF_{Inclusion} = 1.952 * 10^{15}$, highlighting the differential in updating between data and model prediction.

Along with the chronic under-valuing of introduced evidence (in line with other evidential reasoning findings, see e.g., Faigman & Baglioni, 1988; Nance & Morris, 2005), participants consistently under-valued the support provided by dependent cases, irrespective of predicted dependent advantage (e.g., partial and contradicting instances).

Further research We have shown that people under-value the import of dependent reports, even when by their own lights these dependencies yield an informational advantage. However, the question remains open as to whether such a bias can be overruled. Such a question is also worth extending to other forms of structural dependencies, such as "shared backgrounds" (Bovens & Hartmann, 2003; Hahn et al., 2016).

One important caveat to the present work is whether people's down-weighting of dependent evidence is a consequence of poor combination of uncertain evidence (in other words, a computational failure), or due to them having an alternative structural representation of the dependency.

Given that the issue of dependent evidence is so fundamental to many professions and everyday life, it is important to explore these matters further.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research projects Activity (IARPA), under Contract [2017-16122000003]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Open Practices

All data and materials have been made publicly available via the Open Science Framework at (https://osf.io/c7m49/).

References

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

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, 5, 1-17.

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.

Hahn, U., Harris, A. J., & Corner, A. (2016). Public reception of climate science: Coherence, reliability, and independence. *Topics in cognitive science*, *8*(1), 180-195.

Hahn, U., & Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review*, 114, 704–732.

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.

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 (2017). JASP (Version 0.8.4)[Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

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.

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

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.

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.

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

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