Improving the production and evaluation of structural models using a Delphi process

Fergus Bolger[1], Erik P. Nyberg[2], Ian Belton[3], Megan M. Crawford[3], Iain Hamlin[3], Ann Nicholson[2], Abraham Oshni Alvandi[2], Ross Pearson[2], Jeff Riley[2], Aileen Sissons[3], Courtney Taylor Brown Lūka[3], Shreshth Thakur[2], Alexandrina Vasilichi[3] & George Wright[3,4]

Abstract

Bayes Nets (BNs) are extremely useful for causal and probabilistic modelling in many real-world applications, often built with information elicited from groups of domain experts. But their potential for reasoning and decision support has been limited by two major factors: the need for significant normative knowledge, and the lack of any validated methods or software supporting collaboration. Consequently, we have developed a web-based structured technique – Bayesian Argumentation via Delphi (BARD) – to enable groups of domain experts to receive minimal normative training and then collaborate effectively to produce high-quality BNs. BARD harnesses multiple perspectives on a problem, while minimising biases manifest in freely interacting groups, via a Delphi process: solutions are first produced individually, then shared, followed by an opportunity for individuals to revise their solutions. To test the hypothesis that BNs improve due to Delphi, we conducted an experiment whereby individuals with a little BN training and practice produced structural models using BARD for two Bayesian reasoning problems. Participants then received 6 other structural models for each problem, rated their quality on a 7-point scale, and revised their own models if they wished. Both top-rated and revised models were on average significantly better quality (scored against a gold-standard) than the initial models, with large and medium effect sizes. We conclude that Delphi – and BARD – improves the quality of BNs produced by groups. Further, although rating cannot create new models, rating seems quicker and easier than revision and yielded significantly better models – so, we suggest efficient BN amalgamation should include both.

Keywords: Bayes Nets, causal models, crowdsourcing, aggregation, group processes, Delphi

Acknowledgement: This research is based on 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.

[1] Corresponding author, Strathclyde University, Glasgow, UK fbolger42@gmail.com

[2] Monash University, Melbourne, AU

[3] Strathclyde University, UK

[4] **Author note**: The experiment reported here is part of a large project with many personnel. All authors listed contributed to the paper in some way or another e.g. developing the BARD system and training, constructing experimental materials, running participants, analysing data etc. The first two authors were principally responsible for the design of the study and the writing of this report, the remaining authors are listed in alphabetical order.

1 Introduction

1.1 Psychology of probability and causation

Decisions under uncertainty are ubiquitous and difficult. Here, probability and causation are vital elements for accurately representing the world: causal connections constrain what could be influenced by each action, and probabilities quantify our credence in each possible effect. Hence, Bayesian updating and causal decision theory are now central, widely accepted ideals for reasoning in both cognitive psychology and philosophy: they subsume as special cases both the study of logical fallacies and reasoning schemes (Korb, 2003; Hahn & Hornikx, 2016) and philosophical accounts of scientific reasoning (Howson & Urbach, 2006; Corner & Hahn, 2009), and provide general normative standards for measuring reasoning performance in all sorts of domains.

Unfortunately, people have serious cognitive limitations in these areas, and struggle with confusion, complexity, and calculation. Even in simple cases, people are prone to fundamental confusions about elementary Bayesian analysis: how to identify and combine coherent priors and likelihoods to produce posteriors. Despite some recent reconsideration of the evidence for various reasoning fallacies (Hahn & Harris, 2014), this weakness still seems evident in the neglect of base rates, i.e. inadequately weighting the priors (Kahneman, Slovic & Tversky, 1982), in the confusion of the inverse, i.e. interpreting the likelihood as a posterior (Villejoubert & Mandel, 2002), and in the conjunction fallacy, i.e. assigning a lower probability to a more general outcome than to one of the specific outcomes it includes (Jarvstad & Hahn, 2011), as well as more domain-specific versions of these general fallacies, such as the prosecutor’s fallacy (Fenton & Neil, 2000) and the jury observation fallacy (Fenton & Neil, 2000). Similarly, the well-known trap of confusing correlation with causation repeatedly snares even good research scientists (some examples are discussed in Korb, Hope & Nyberg, 2009 and Lawlor, Davey Smith & Ebrahim, 2004), which may well lead to overestimating the likely impact of policy actions. Common interrelationships between several variables can easily result in misunderstanding the significance of evidence, e.g. screening off (Korb & Nicholson, 2011), explaining away (Liefgreen, Tešić, Lagnado, 2018) and the zero-sum fallacy (Pilditch, Fenton & Lagnado, 2018). Finally, as the number of variables and nonlinear connections grows even larger, the number of conditional probabilities involved increases super-exponentially (Korb & Nicholson, 2011), so mentally computing them becomes impossible.

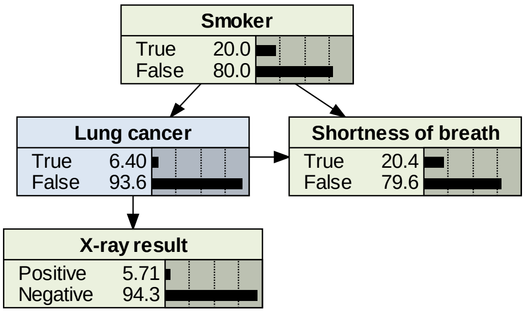
1.2 Causal BNs

Fortunately, we have an Artificial Intelligence (AI) technology that models such situations, and performs probabilistic and causal reasoning: Bayes Nets (BNs). Technically, a BN is a directed, acyclic graph whose nodes represent random variables, and whose arrows represent direct probability dependencies quantified by conditional probability functions associated with each node. For example, in Figure 1(a), the *Lung cancer* node represents whether a patient has lung cancer (a ‘discrete’ variable, i.e. it has a finite number of possible states). The arrow from the *Smoker* variable shows that the probability of *Lung cancer* = *True* is specified (using the conditional probability table (CPT) shown) dependent on one condition: whether the patient is a smoker. In a causal BN, such as this one, the arrows also represent causal influence, e.g. smoking directly causes the increased probability of lung cancer. As shown in Figure 1(b), users can enter evidence about any variables, e.g. *Smoker* = *True* or *X-ray result* = *Positive*, which is efficiently propagated in standard software packages to update the probability distributions for all other variables. Thus, causal BNs can support and perform diagnostic, predictive and decision-oriented probabilistic reasoning. Furthermore, users can perform sensitivity analysis (e.g. what if the evidence or model were slightly different?), and re-use previous BNs (e.g. revising or expanding models to improve performance or apply them to similar domains). For further detail, see (Spirtes, Glymour & Scheines, 2000; Korb & Nicholson, 2011).

BNs are a general-purpose technology and have already been deployed for many purposes in diverse domains, notably medicine (e.g. Flores et al., 2011; Sesen et al., 2013), the environment (e.g. (Wintle & Nicholson, 2014; Chee, 2016), and the law (e.g. Fenton & Neil, 2000; Fenton, Neil & Lagnado, 2013; Neil et al., 2019). The US Intelligence Advanced Research Projects Activity (IARPA) have also shown interest in having groups of intelligence analysts build relevant causal BNs to support their analysis, and hence subsequent decision and policy making, by funding our current research under the CREATE program (Crowdsourcing Evidence, Argumentation, Thinking and Evaluation, see www.iarpa.gov/index.php/research-programs/create).

Researchers have developed machine learning methods to learn causal BNs from data (summarised in Korb & Nicholson, 2011), such as learning the BN in Figure 1 from a medical database. Psychological methods have also been developed to elicit the necessary information from domain experts, which can then be used by a BN expert to build a good model (as described in Korb & Nicholson, 2011). More recently, techniques have been developed to help combine them (e.g. Flores et al., 2011). Provided an accurate causal BN has been built, then the reasoning it produces will be accurate, and avoid most of the difficulties and fallacies that bedevil humans. For instance, Figure 1(a) illustrates that the probability of lung cancer depends upon the probability that the individual is a smoker, which might be estimated from the base rate of smoking in the relevant population. Unlike many humans, the BN will not neglect the base rate in computing the consequent probability of lung cancer. Furthermore, when presented with the BN model to view and explore, human users are more likely to understand why the base rate is relevant. More generally, BNs have great potential as ‘explainable AI’ that can be understood, used and trusted by ordinary people, because the nodes and arrows of BNs are designed to have clear semantic interpretations – unlike the ‘black boxes’ provided by artificial neural nets (Bostrom & Yudkowsky, 2014).

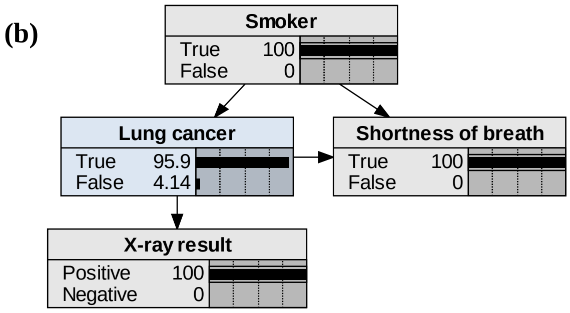
(a)



“In the absence of evidence, the prior probability of *Lung Cancer=True* is 6.4% (almost certainly not).”

Smoker	Lung cancer	
	True	False
True	30%	70%
False	0.5%	99.5%

Lung cancer	X-ray result	
	Pos	Neg
True	60%	40%
False	2%	98%



"Observing *Shortness of breath = True*, *X-ray result = Positive* and *Smoker = True* increases the probability of *Lung cancer = True* to 95.9% (almost certain)."

Figure 1: A simple medical diagnosis BN (a) without any evidence (also showing two of the CPTs), and (b) with three pieces of evidence added. Quoted underneath is some of the automated verbal explanation provided by the BARD BN software.

Where an accurate BN isn't already available, there are good theoretical reasons to think that *the process of constructing one* will help to avoid many of the cognitive limitations listed above (Korb, 2003; Korb & Nyberg, 2016). In this example, provided that a domain expert knows that smoking affects the probability of lung cancer and has correctly placed an arrow from the former to the latter, then there are immediately empty boxes in the CPT for *Smoker* corresponding to its base rate that must be filled in – so these base rates cannot be neglected. Empirical results from our other CREATE experiments support this general thesis. For example, the following reasoning difficulties were overcome more often by individuals and/or groups able to construct the relevant BN than by those given generic critical thinking advice: the zero-sum fallacy (Pilditch, Fenton & Lagnado, 2018), integrating dependent evidence (Pilditch, Hahn & Lagnado, 2018; Korb et al., 2019) ; explaining away (Liefgreen, Tešić, Lagnado, 2018; Korb et al., 2019), and duplicitous witnesses (Pilditch, Fries & Lagnado, 2019).

One major barrier to using BNs for reasoning and decision support has been the need for significant normative knowledge to understand and build them. Even where knowledge is elicited from domain experts, BN experts are usually required to construct the network. For domain experts to construct their own BNs, substantial training is usually needed. The major software packages have very limited integrated help or training, both in scope and presentation. Hence, typically, an introductory course in BN modelling takes 2–3 days of upfront instruction from an expert (e.g. two days by Bayesian Intelligence, see bayesian-intelligence.com; three days by BayesiaLab, see www.bayesia.com). While many domain experts have made this commitment and successfully used BN modelling in their work, many others have balked at it. IARPA's informal survey of intelligence analysts for CREATE, for example, revealed the maximum most analysts would commit to upfront training was 0.5–1 day. A second barrier is the lack of validated methods and software support for groups to build BNs collaboratively, including the amalgamation of diverse opinions into a single group product. This is particularly important if domain experts are trying to pool their own wisdom (including their limited modelling skills).

Our aim in CREATE was to design, build and validate such a method, software platform and training: enabling domain experts to undergo, at most, 0.5 days of upfront training, then exchange information easily – and use an effective procedure for combining that information – to create appropriate BN models to assist with subsequent written reports (further details in §1.4).

1.3 Delphi Processes for Groups

Groups (e.g. Hackman & Katz, 2010; Hastie & Kameda, 2005), or the wider 'crowd' (e.g. Brabham, 2008; Surowiecki, 2005), have been shown to outperform individuals on judgment tasks, presumably by bringing together multiple perspectives and through the exchange of heterogeneous knowledge. There is also some evidence that individuals within groups may be motivated to work harder than alone (e.g. Weber & Hertel, 2007) – a process generally known as social facilitation (Zajonc, 1965).

The advantage gained by taking part in a group is often termed 'process gain'. However, if the group interactions are not managed properly, then their potential advantages may not be realised – because groups may also suffer from processes that can undermine their effectiveness, leading to 'process loss' (Steiner, 1972). In fact, the operation of well-documented social biases may lead to inferior outcomes from groups relative to individuals. For example, individuals in groups do not always share their unique information, focusing instead on the information that is commonly held by group members (e.g., Stasser & Vaughan, 2013). Further, people in groups may suffer social inhibition instead of facilitation – due, for example, to anxieties related to interacting with strangers (Buck et al., 1992) – reducing their motivation to perform: social facilitation has been found to turn into inhibition as task complexity increases, perhaps due to corresponding increases in levels of arousal (Zajonc, 1980) or evaluation apprehension (e.g., Blascovich et al., 1999). There can also be a diffusion of responsibility in groups leading to reduced participation, sometimes referred to as social loafing: the probability of such behaviour increases with group size and degree of physical and temporal dispersion of group members (Chidambaram & Tung, 2005). Further, a number of well-documented cognitive biases reduce the quality of individual judgment and reasoning (e.g., Arnott, 2006; Kahneman, Slovic & Tversky, 1982; Morvan & Jenkins, 2017) and these might be exacerbated by group processes. For example, individuals have often been found to be overconfident (e.g., Johnson & Fowler, 2011; Lichtenstein, Fischhoff & Phillips, 1982), and groups demonstrate shifts towards even greater risk taking and confidence in judgment than by their component members (Dodoiu, Leenders & van Dijk, 2016; Stoner, 1968).

Various group structured techniques have been developed to help improve the outcomes produced by freely-interacting groups: one of the best-known of these is the Delphi technique (Linstone & Turoff, 1975). Delphi is an example of a ‘nominal-group’ technique, where the group members never actually meet face-to-face – and are, in fact, anonymous to each other – interacting only through a facilitator. This reduces the undue influence of more powerful or dogmatic individuals, so participants can focus on providing and assessing judgements and reasoning more frankly. In the first round, each participant must first make a judgment individually (before seeing any other responses) and submit it to the facilitator, perhaps along with their confidence in and/or a rationale for it. This reduces social loafing and premature conformity (as seen in anchoring and groupthink), thus increasing the independence and diversity of initial responses.

The facilitator then provides all the participants with feedback, thereby initiating the second round. For quantitative estimates, this usually includes some statistical summary such as their mean and standard deviation, and often the individual estimates with associated confidence and/or rationales if they were collected. Participants are invited to revise their responses if they wish and resubmit; this is the end of the second round. Anonymity reduces the social pressure to conform, but feedback encourages participants to rationally reconsider their response in the light of the new information provided.

The facilitator can always provide feedback again, thus initiating another round, if any significant revision seems likely – but two or three rounds are usually sufficient. This process tends to increase the level of consensus in the group, but the more fundamental aim is to increase the overall quality of the responses. Using a facilitator, in addition to aiding administration and collation, tends to encourage constructive contributions from members and avoid any unproductive, heated arguments.

After the final round, the facilitator usually aggregates the responses of the individual members (or collates them, if they are qualitative), and the result is taken as the group response. Each opinion is given equal weight throughout, unless there is some strong, evidence-based reason for valuing participants' opinions differently. The process does implicitly 'weigh' answers in another way, which may well be encouraged (Bolger & Rowe, 2015): those who are more sure of their judgments – after considering the opinions of others – should stick to them, whereas those who are less sure should change. This is in line with the Theory of Errors (Dalkey, 1975; Parenté & Anderson, 1987), which proposes that Delphi produces virtuous opinion change by means of the less knowledgeable being persuaded towards the position of the more knowledgeable.

In summary, the four main features of Delphi that help reduce process loss (Rowe, Wright & Bolger, 1991) are: anonymity, iteration, feedback, and aggregation. A review by Rowe & Wright (1999) found that, at least for short-term forecasting problems and tasks involving judgements of quantities, Delphi has generally shown improved performance compared to freely interacting groups or a statistically aggregated response based on the first-round responses of individual participants.

1.4 Delphi Groups for Causal BNs

Previously, the Delphi technique has mainly been limited to either making relatively simple judgements (e.g. forecasts of uncertain quantities), or answering rather specific questions in a qualitative mode (so-called 'policy Delphi' (Turoff, 1970). Recently, one study used a form of Delphi for point-estimate CPT elicitation (Etminani, Naghibzadeh & Peña, 2013), while for BN structure elicitation Serwylo (2015) pioneered online crowdsourcing and automated aggregation (albeit non-Delphi). We further developed these ideas and built a GUI to automate both structure and parameter elicitation and amalgamation (Nicholson et al., 2016). However, our current CREATE research is the first project to apply Delphi to developing and exploring an entire BN model – including variables, structure and parameters – and also to complex reasoning problems.

To overcome the existing barriers to using BNs, we developed a computer-based structured technique – Bayesian Argumentation via Delphi (BARD) – to permit groups of domain experts with minimal normative training to produce high quality BNs. Here, we only explain how this experiment (the specific task for participants and the social aggregation process) fits into the overall BARD system; for more detail on BARD see (Nicholson et al., 2019).

1. To reduce the training barrier, BARD requires only 2.5 hours of upfront training, which is provided online as an e-course. This and other training resources can subsequently be accessed on demand within the BARD platform, in addition to context-sensitive tips and help. Our experimental subjects had undergone this training as part of a previous experiment.
2. To facilitate group interaction, BARD is provided on a web-based platform accessible from anywhere at any time, by any number of individuals. Each member has their own virtual workspace that can be shared and compared with collaborators, in addition to a group workspace for displaying the consensus model.
3. To assist both inexperienced users and aggregation, BARD decomposes the task into a logical series of smaller steps to construct BNs and produce reports. Most relevantly for this experiment, variables are specified at Step 2, arrows between them at Step 3, and probabilities at Step 4.
4. To aggregate individual responses, BARD uses a Delphi-like process at each Step. The first time each user encounters each Step, they must attempt to answer themselves. Then they are able to see any other answers others have already given, at which point they can revise and resubmit their answer, and also (at least for some Steps) rate other users' responses. This process may be moderated by a facilitator with no special domain or BN expertise, or automatic algorithms may be applied to manage participants and aggregate responses (using response frequencies and/or ratings).

In previous studies, we showed that individuals with BARD (i.e. without the usual group interaction, but using the BARD training, software and approach to construct appropriate BNs) could solve some specific causal, probabilistic reasoning problems (which included known fallacies) far better than control individuals without BARD (Liefgreen, Tešić, Lagnado, 2018; Pilditch, Fenton & Lagnado, 2018; Pilditch, Hahn & Lagnado, 2018). In a subsequent experiment, we also showed that groups with BARD performed far better than control individuals without BARD (Korb et al., 2019). In the current experiment, we complete this three-way comparison: we examine whether groups with BARD outperform control individuals with BARD, or (more specifically) whether the Delphi-like social processes within BARD help to improve group solutions after Round 2 beyond those of BARD individuals after Round 1. We are not, here, comparing the quality of solutions to control individuals without BARD.

1.5 Hypotheses

In this experiment, we only investigated the effectiveness of a Delphi-like process on the production of structural models, i.e. where participants select appropriate variables and arrows and compare work without having entered probabilities, which is consistent with the overall BARD process. We were not investigating other aspects of BARD, such as variable generation, parameter estimation, or report writing. As such, we just used a part of BARD (Steps 2 & 3) and only used problems amenable to causal BN modelling.

Our research question was simply: ‘Will there be a significant improvement in the quality of structural models produced if there is an opportunity by analysts to see several other models of varying quality in addition to their own?’ This is operationalised as a test of a single hypothesis with two parts. After analysts have had the opportunity to review several diverse models produced by peers in Round 1:

H1(a): The average quality of the structural model most highly rated by each analyst at Round 2 will be higher than the average quality of each analyst’s individually-produced model at Round 1.

H1(b): The average quality of the structural model produced by each analyst will increase due to their revisions in Round 2.

2 Method

2.1 Performance Measure

Rather than asking participants to define variables from scratch, we offered them a set of predefined, plausible variables from which to choose. This had two important advantages:

1. It made the task easier and quicker for participants, both because they did not need to define

variables in Round 1, and because they could more easily compare their work with that of their peers in Round 2. This minimised dropout and allowed them to solve two test problems within a two-hour session.

2. It simplified the assessment of answers, both because it avoided minor semantic variations in the definitions of variables, which would have required human judgement to assess as equivalent, and because it avoided more significant variations that would allow alternative graph structures to be acceptable answers (e.g. merging two binary variables into one 4-state variable). Given the nature of the problem statement, some variables in our set were well-defined, individually necessary, and collectively sufficient, whereas the remainder were unnecessary and/or badly defined – although tempting enough to induce some modelling mistakes. Thus, all our BN experts agreed there was a clear, uniquely best answer to each problem, which we will refer to as the 'gold-standard' model (for further details, see the experimental materials provided in our linked Supplementary Materials).

All the participant models could then be scored against the gold-standard model using arrow edit distance (AED), which simply counts how many arrow 'edits' (additions, deletions, and reversals) are required to convert one BN into the other. This is far and away the most common such measure in the relevant literature (e.g., Spirtes, Glymour & Scheines, 2000).

Despite its popularity, we note that AED has clear limitations: it only measures some of the errors that might be made in constructing a BN (limited scope), and it gives independent and equal weight to each of the errors (linearity). Regarding scope, although participants in our experiment notionally needed to choose which variables to include in their models as well as how to connect those variables with arrows, AED was nevertheless suitable. This is because (i) participants could not make any errors within the definitions of variables, so these did not need be assessed, and (ii) including (or excluding) a variable is equivalent to having (or not having) at least one arrow that connects this variable to others, so any error in the variables chosen must translate into at least one associated arrow edit. Regarding linearity, future research could certainly develop and justify more complex measures for specific purposes. These might take into account the frequency with which particular errors are made, their adverse consequences for the model, or the underlying cognitive processes responsible for the error. However, absent such research, it is reasonable to use AED here as the most widely accepted and general-purpose measure of performance.

Using predefined variables and AED also had the advantage that we were able to develop software scripts to automatically assess the BNs submitted by participants, rather than employing human raters. Hence, no blinding of raters was required to the Round in which models were produced.

2.2 Participants

In a previous experiment using BARD (described in Korb et al., 2019), 145 Monash University students received the 2.5 hours of BARD training, and in their assigned groups of 6–8 people, attempted three problems requiring complete BN models as solutions. From these earlier participants, we enlisted a sample of 98 for this experiment – the maximum we were able to obtain – to attend one of three sessions on consecutive days. For logistical reasons, it was only possible to run a maximum of 25 participants in each session, so several participants from the pool were unable to find a convenient slot and did not sign up. With no shows, and a few participants failing to complete – or properly complete – the tasks within the allotted time (2 hours, which surveys indicated was adequate for the vast majority of participants), we were left with 57 usable data sets. Power analysis showed that 52 analysts gives a power of 0.8 to detect the expected medium effect size of 0.35 (typical of our pilot studies). We collected data from all analysts who attended, and stopped collecting data after the three sessions were complete.

Participants were approached after they had completed the earlier BARD study: they were thus self-selected from these ‘BARD graduates’. They were offered an A$50 gift certificate for two hours’ work. Subsequently, after they had consented to participate in the experiment but before they began solving the problems, they were told that for each perfect BN answer (i.e. gold-standard model) – which they could submit in the first and/or second round, and for the first and/or second problem, hence a maximum of four per person – each participant in the experiment would be awarded a ticket in a single lottery for an additional A$200 gift certificate. This lottery was conducted under the joint supervision of the experimental leads, and the winner notified by email.

2.3 Design and Materials

The experiment is an entirely within-subjects design: each participant created a model individually for each of two problems, and the quality of these models was compared to both the quality of the peer model for each problem that the participant rated most highly, and their own revised models after receipt of feedback. Confidence intervals were computed separately for the two problems. Thus, each problem is properly considered a replication to explore the generalizability of findings.

Our other BARD team members at University College London (UCL) and Birkbeck – both of the University of London – had previously developed suitable test problems. These are relatively simple Bayesian reasoning problems, but incorporate some tempting qualitative fallacies, so building an appropriate BN achieves quantitative accuracy and avoids these qualitative traps. The two problems we used are called Black Site, which incorporates the zero-sum fallacy (Pilditch, Fenton & Lagnado, 2018), and Spider, which incorporates potentially duplicitous witnesses (Pilditch, Fries & Lagnado, 2019); these materials are provided in our Supplementary Materials.

Using the Simulated Group Response Paradigm (SGRP, described in Bolger et al., 2020), we first conducted piloting in which participants who had already undergone BARD training and produced BNs using BARD were presented with new problems requiring structural solutions. We then created two alternative stimulus sets of 6 models for each problem, based on those elicited during piloting. These sets were designed to be representative of the types of models produced by trained participants, or which might well be produced (e.g. including a type of error that might be expected to occur but was not observed in the pilot responses). Each set was composed of solutions of varying types and correctness in roughly the proportions observed in piloting; for Black Site, this resulted in four gold-standard and two erroneous models in each of the two stimulus sets, and for Spider, three gold-standard and three erroneous models in each set – see our Supplementary Materials for details[5]. In the main study, new participants – who had also undergone training in BARD – were randomly assigned to each of the two

[5] Piloting used two batches of participants: stimulus sets were developed based on participant responses in an earlier BARD experiment at UCL, then refined based on a smaller batch of responses from our Monash recruits. We did not run a formal Stimulus Elicitation Exercise (SEE) as is common for the SGRP, but instead used 'representative selection' (see Bolger et al., 2019 for details). Due to the relatively small sample sizes of trained analysts solving each problem during the piloting, random sampling from a SEE database would probably not produce representative stimuli sets (and the piloting may well not produce any examples of some plausible categories of response).

sets in approximately equal numbers, and presented with these solutions in Round 2 as if they were responses by other members of a Delphi group working on the same problem (there was actually no deception here: participants were simply told that they would see some other structural models that were offered as solutions to the problem).

2.4 Procedure

75 different people to those participating in the pilot studies were signed up to one of three experimental sessions. After receiving instructions and providing consent, each participant took part in two successive rounds for the first problem: 15 minutes for Round 1, a 5-minute break, and 20 minutes for Round 2. After another short break, they took part in two similar rounds for the second problem. Finally, they completed a brief survey asking whether, for each of these four rounds, they had enough time to complete the task.

Round 1: Each analyst received a description of a problem they had not encountered before, plus associated lists of variables. They were asked to independently produce an appropriate BN structure for the problem by choosing variables and connecting them with arrows.

Round 2: The analyst saw 6 other models for the problem (one of two possible stimulus sets). They were asked to (a) rate the quality of all 6 models (and their own), and (b) improve their own model (and self-rating) if possible.

Analyst ratings were on a 7-point Likert scale with two poles: ‘poor’ to ‘excellent’, with the intermediate response options not labelled. This corresponds to the rating system used in BARD. After all experimental sessions had been completed, all participants were informed via email of the gold-standard structures for these problems, and were provided with their compensation. Subsequently, the quality (i.e. the edit distance from the gold-standard) of each analyst’s Round 1 model was compared to (a) the quality of the model they rated most highly at Round 2 (excluding their own), and (b) the quality of the analyst’s revised model after Round 2.

3 Results

3.1. Confirmatory analyses

Table 1 shows the sample mean edit distances (and SDs among participants) by problem and source of model. Note that smaller edit distances indicate better BNs (i.e. more similar to the gold standard). 95% two-sided CIs for the population means are shown visually in Figure 2.

Problem	*Individual BN*	*Top-rated BN*	*Revised BN*
Black Site	7.05 (4.28)	1.47 (2.56)	3.86 (4.64)
Spider	5.32 (3.87)	1.13 (1.59)	3.54 (4.46)

Table 1: AED means (and individuals' SD) by problem and source of model.

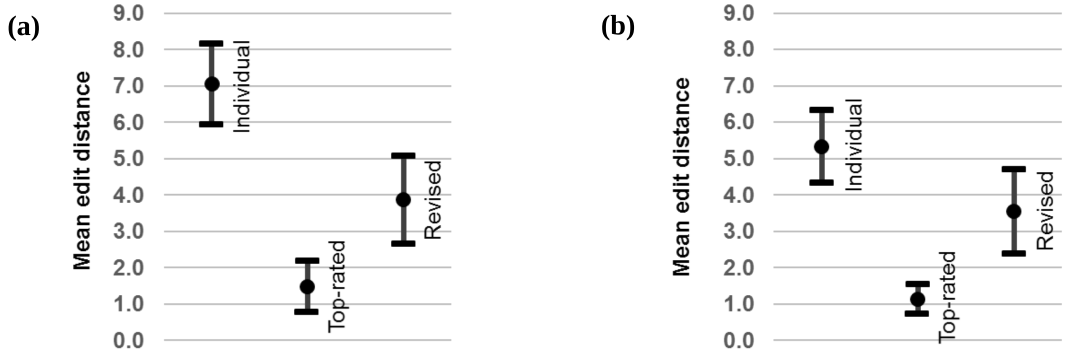


Figure 2: 95% CIs for mean quality of models for (a) Black Site, and (b) Spider.

A paired, two-sided *t*-test was performed to test H1(a) and H1(b) for each of the two problems. The variables used for these tests were the quality of each participant's Round 1 and Round 2 structural model, where the Round 1 was individually produced and the Round 2 model is either the revised model after feedback (H1(b)) or the highest rated of the other models presented excluding the participant's own (H1(a)). Therefore, there were four tests in total – the interpretation of the results take into account the slightly increased family-wise error rate that this procedure entails.

For each effect size, we computed both the two-sided 80% and 95% CIs for the mean differences in raw AED score. We also standardised these effect sizes via Glass's **Δ**, i.e. expressing them as fractions of the standard deviation of individuals' Round 1 scores.[6] If the two-sided CI for a difference in scores does not overlap 0, then this is mathematically equivalent to rejection of the null hypothesis by our corresponding two-sided *t*-test.[7]

Any participant who failed to submit all his/her revised models was to be excluded from the

[6] Glass's Δ is a more appropriate variation of Cohen's *d* designed for experiments in which the variance within the two groups is unequal (Glass, Peckham, & Sanders, 1972), which we correctly anticipated here, since in a Delphi process variance tends to decrease as participant answers converge.

[7] Notwithstanding this equivalence, we are contractually obliged to base our inferences primarily on such CIs and standardised effect sizes, rather than the null-hypothesis significance tests *per se*.

analysis, but this did not occur. However, four participants made no or partial ratings for either problem and ten participants provided no or partial ratings for one of the problems; these participants were excluded from the analysis of ratings for the problem concerned.

Test of H1(a): The hypothesis was supported for both problems with large and very large effect sizes; these results are shown visually in Figure 3(a).

Black Site: $\mu_{AED1} - \mu_{AED2a}$ = 5.59 fewer edits (95% CI [4.30, 6.88], 80% CI [4.75, 6.43])

Glass's Δ = 1.27 (95% CI [0.78, 1.76], 80% CI [0.95, 1.60])

Spider: $\mu_{AED1} - \mu_{AED2a}$ = 3.78 fewer edits (95% CI [2.80, 4.77], 80% CI [3.15, 4.42])

Glass's Δ = 1.05 (95% CI [0.58, 1.51], 80% CI [0.74, 1.35])

Test of H1(b): The hypothesis was supported for both problems with medium and large effect sizes; these results are shown visually in Figure 3(b).

Black Site: $\mu_{AED1} - \mu_{AED2b}$ = 3.19 fewer edits (95% CI [1.84, 4.55], 80% CI [2.32, 4.07])

Glass's Δ = 0.76 (95% CI [0.36, 1.15], 80% CI [0.50, 1.01])

Spider: $\mu_{AED1} - \mu_{AED2b}$ = 1.77 fewer edits (95% CI [0.84, 2.71], 80% CI [1.17, 2.38])

Glass's Δ = 0.45 (95% CI [0.07, 0.83], 80% CI [0.20, 0.70])

As we noted above, repeating tests several times increases the family-wise error rate. To take this into account, we repeated tests of H1(a) and (b) with a very strict criterion (CI = 99.9%), but the CIs still did not cross zero, i.e. the hypotheses were still supported.

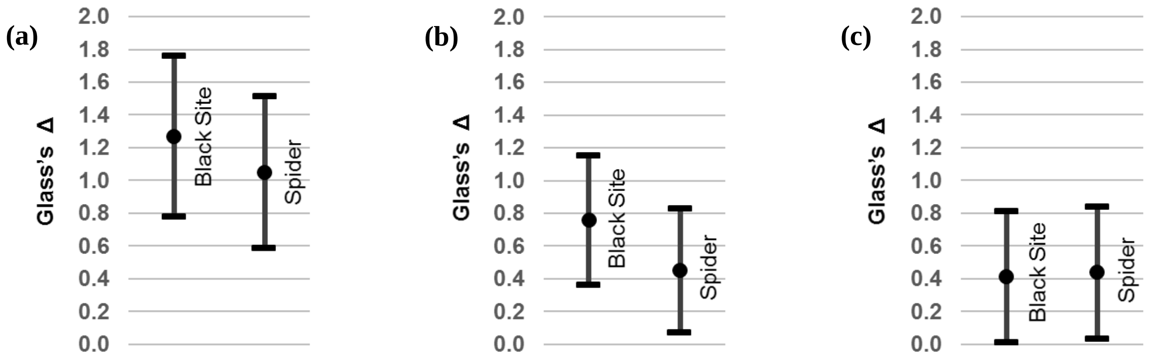


Figure 3: 95% CIs for quality differences between models, showing:
(a) individual > top-rated, (b) individual > revised, and (c) top-rated > revised.

3.2. Exploratory Analyses

There are many further interesting questions about participant performance, and although the dataset is limited, we performed exploratory analyses to obtain some indication of the answers and guide future research. We summarise the findings here:

1. How many Round 2 answers were better, worse, or of the same quality as the participant's Round 1 answer?

Pooling results from the two problems (i.e. each participant producing two models), in 63 cases (56%) the participant improved their answer, which was nearly five times more common than the 13 cases (12%) where the participant made their answers worse, while in 36 cases (32%) the participant submitted the same quality answer. Thus, the improvement due to Delphi was relatively widespread among participants, rather than due to just a few giving very poor answers in Round 1 and then making large improvements in Round 2. This suggests the Delphi effect will be robust.

2. If a participant makes more mistakes, then are they more or less likely to improve their answers?

If a participant makes more mistakes, then this is evidence that they have less modelling ability (at least for these test problems). So, one might expect them to have less ability to judge which models are better (addressed in question 4) – and more specifically, less ability to recognise that an alternative model is better where they have made a mistake, resulting in fewer corrections. This would run counter to the Theory of Errors and reduce the effectiveness of Delphi for BNs: bad models would improve least, and the spread between good and bad models would increase.

Spearman's rho (r_s) was used to compute the rank correlation between the number of errors made by a participant and two different measures of improvement in their answers (N=57 for both). Participants who gave the gold-standard answer in Round 1 are included here even though they could not improve (and might do worse) in Round 2, but it makes little difference if they are excluded.

Firstly, when measuring change in terms of the nett *number* of mistakes corrected, there was a significant positive correlation for both test problems:

Black Site: AED_1 correlated to $\mathrm{AED}_1 - \mathrm{AED}_{2b}$ ($r_s = 0.42$, 95% CI = [0.66, 0.14]).

Spider: AED_1 correlated to $\mathrm{AED}_1 - \mathrm{AED}_{2b}$ ($r_s = 0.35$, 95% CI = [0.59, 0.08]).

Secondly, when measuring change in terms of the nett *proportion* of mistakes corrected, there was no significant correlation for either test problem:

Black Site: AED_1 not correlated to $(\text{AED}_1 - \text{AED}_{2b})/\text{AED}_1$ ($r_s = 0.08$, 95% CI = [–0.19, 0.34])

Spider: AED_1 not correlated to $(\text{AED}_1 - \text{AED}_{2b})/\text{AED}_1$ ($r_s = 0.12$, 95% CI = [–0.12, 0.37])

Taken together, this suggests that participants who made more mistakes were *not* worse at correcting each of their individual mistakes, so (since time permitted) they made more corrections in total. Thus, not only did participant scores improve overall, but also the variance in participant scores was reduced. These results support and extend the main conclusion that participants with worse BN structures were willing and able to improve them after seeing better BN structures. Also, although Delphi does not rely on convergence in either answers or answer quality (since average answer quality can improve without convergence), the ideal outcome of group collaboration is convergence on the best answer (where that exists, as here), so the convergence is encouraging.

3. Are the solutions produced by the rating method better or worse than those produced by the revision method?

The quality of solutions produced by rating and revision were compared using a paired, two-sided *t*-test and CI on mean difference between top-rated and revised models. Top-rated models were significantly better than revised models, with a medium effect size[8]; these results are shown visually in Figure 3(b).

Black Site: $\mu_{\text{AED2a}} - \mu_{\text{AED2b}}$ = 1.79, 95% CI [0.68, 2.89], 80% CI [1.07, 2.50]

Glass's Δ = 0.41, 95% CI [0.01, 0.81], 80% CI [0.15, 0.67])

Spider: $\mu_{\text{AED2a}} - \mu_{\text{AED2b}}$ = 1.76, 95% CI [0.79, 2.73], 80% CI [1.13, 2.38]

Glass's Δ = 0.44, 95% CI [0.03, 0.84], 80% CI [0.17, 0.70])

This result may seem surprising, since participants had the opportunity to revise their own models to mirror their top-rated stimulus model. However, where a participant's revised model was *worse* than the gold-standard, they may well have rated the gold-standard as the best alternative model, but either (i) mistakenly thought their own model was better, or (ii) not made the additional effort to revise their own. Furthermore, it wasn't possible for a participant's revised model to be *better* than the gold-standard, which was always one of the models they rated – in this respect, experimental conditions were

[8] The exact effect size given here varies slightly from the difference between the effect sizes for H1(a) and H1(b), because the samples used vary slightly: some participants didn't provide ratings, so they were excluded from the calculations here and for H1(a), but were included in the calculation for H1(b).

favourable for the rating method.

4. If a participant provides better answers themselves, then do they also give more accurate ratings for other answers?

As noted, if a participant constructs a better model, then one might expect that they also have more ability to rate other models. To test this, we first measured a participant's rating accuracy for other answers by the rank correlation (Spearman's r_s) between all their ratings of stimulus BNs and the attendant edit distances of those BNs. (It was not possible to create a rating accuracy for all participants, because some participants gave identical ratings to all BNs.) The average rating accuracy was $r_s = 0.34$ (medium) for Black Site and $r_s = 0.6$ (large) for Spider. These good correlations further substantiate our general finding H1(a) that participants can tell which models are better. In turn, the rank correlation between each participant's rating accuracy and their own answer quality in Round 1 was then calculated: $r_s = 0.18$ (small) for Black Site (N=47) and $r_s = 0.05$ (zero) for Spider (N=38). Similarly, the rank correlation between each participant's rating accuracy and their own answer quality in Round 2 was positive but small for both problems.

The surprisingly small correlations between answer quality and rating accuracy have interesting implications for the use of ratings to help aggregate answers. It suggests that it may not be beneficial to discount the ratings of participants whose own answers were poorly rated (or who show some other indication of giving poor answers), since their ratings may nevertheless be of similar accuracy to the ratings produced by those who gave better answers (and the average rating of a model will be more stable if the number of ratings it includes is larger). Also, it suggests that how many participants give a specific answer (its prevalence) and how highly participants rate that answer may be partially independent indicators of answer quality, so it might be best to use both.

5. What were the relative frequencies of these three categories of mistake: addition (adding arrows that should not be there), deletion (leaving out arrows that should be there), and reversal (putting an arrow that should be there the wrong way around)?

Type	*Possible*	*Round 1*	*Round 2*	*Improvement*
Addition	12996	310	205	105
Deletion	798	288	118	170
Reversal	798	79	72	7
All	13794	677	395	282

Table 2: Frequencies of the three types of mistake: addition, deletion, and reversal.

Pooling the data from both test problems, there were roughly the same number of addition errors (310) as deletion errors (288) in Round 1. However, to assess the propensity of participants to make a specific type of error, we also need to calculate the number of possible errors of this type, and hence how many times such errors were avoided. As with many real BNs, the test problems require participants to add a small fraction of the possible arrows (6/132 for Black Site, and 8/110 for Spider). So, the number of possible addition errors (228 per participant) is much larger than the number of possible deletion errors (14 per participant), and although the actual numbers in Round 1 are roughly equal, they represent only 2% of the possible addition errors but 36% of the possible deletion errors. On this measure, the propensity to wrongly add an arrow is much less than the propensity to wrongly omit one.

There is a plausible two-factor explanation for this disparity: whether a participant makes an arrow error depends on both attention and judgement. Furthermore, addition requires committing a positive action, whereas so-called deletion (i.e. omission) does not. Thus, to commit an addition error, a participant must have attended to the possible arrow but made an error of judgement. In contrast, a participant can make a deletion error either for this reason (with, plausibly, a similar chance of judging incorrectly) or because they did not give adequate attention to the arrow (and it is omitted by default). Since there are many possible arrows to consider in the 15 minutes available for each problem in Round 1, we can expect that many of these arrows were not given much attention – resulting in the higher propensity for deletion errors.

In Round 2, correction rates showed the opposite disparity: the correction rate for addition errors was roughly half as great (34%) as for deletion errors (59%). This is also explained by the two-factor account. To correct an addition error, a participant must amend their original judgement. In contrast, a participant can either correct a deletion error either for this reason (with, plausibly, a similar chance of amendment), or they can belatedly attend to an arrow they previously neglected and decide to include it. When participants are presented with several gold-standard answers, this should draw explicit attention to any differences from their own model – including any gold-standard arrows they neglected – resulting in the observed higher propensity to correct deletion errors than addition errors. In future BARD development and experiments, we may benefit from distinguishing between these two factors, e.g. to increase the adoption of neglected ideas it may be sufficient to increase their salience, whereas to increase the correction of mistaken ideas it may be necessary to facilitate exchange of the reasons that support them.

The number of reversals in Round 1 (79) was comparatively low, at about a quarter of addition or deletion errors, and this represents only 10% of the possible reversal errors. This is encouraging, since a fundamental and frequent misunderstanding about causal BNs is to not appreciate that the arrows must be oriented in causal directions regardless of the direction of evidential inference – so BARD training on this point was evidently effective. However, in Round 2 the nett correction rate for reversals was also very low (9%). Closer inspection showed that there were more corrections to Round 1 reversals than this suggests – but they were matched by new reversal errors introduced in Round 2, often by different participants. Thus, there was a low but persistent base rate for this error, which has several implications. Upfront training on this issue could be improved further, possibly including a hurdle test (i.e. that must be passed to complete the training). The real-time tips could be improved, e.g. if one participant has oriented an arrow (or multiple arrows) in the opposite direction to most other participants, then they could be automatically asked if the arrow(s) are oriented causally. This issue also illustrates a limitation of AED if it is interpreted as a measure of participant understanding. Reversal errors were more likely to be made by participants who made other reversal errors (clustering), because any such error reflects a common cause: a more abstract misunderstanding present in these participants.

6. Which specific errors were the most frequent?

The kinds of errors people often make in BN building have been noted anecdotally in BN textbooks and training, but there has been little formal empirical investigation. Furthermore, applying a Delphi-like process to BN building (where there is an explicit opportunity for errors to be corrected) is novel. So, exploring the kinds of errors made and corrected in our experiment is of great interest to improving our cognitive models of participants and improving our system to avoid such errors (through training, intelligent help, etc). We provide a few examples here to illustrate the methodology.

To summarise these statistics in a readily accessible way, we developed a new type of visualisation, examples of which are shown in Figures 4–6 for the Spider problem. Figure 4 shows that, in addition to the target variable *Spider in facility* that could either be true or false, participants incorrectly included – using similar arrows – *Spider not in facility*, which was both logically unnecessary and had been defined with nonsensical states (see Supplementary Materials).

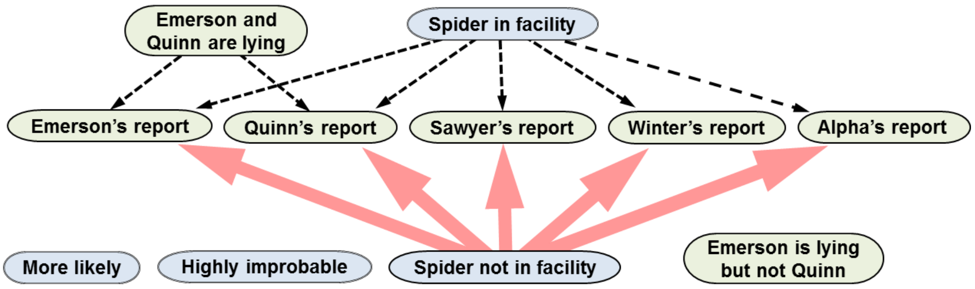


Figure 4: Spider addition errors (solid arrows) and their relative frequency (via thickness), with gold-standard arrows (dashed) for context.

Figure 5 shows that participants had less difficulty in including any of the direct connections between evidence and hypothesis variables than they did in including the connections to the 'background' variable *Emerson and Quinn are lying*, which was neither evidence nor hypothesis but crucial for determining the relevance of the testimony provided by Emerson and Quinn. These two most frequent deletion errors correspond to not properly accounting for the possibility – explicitly discussed in the problem statement, but initially unlikely – that the two sources are duplicitous conspirators. Figure 6 shows that this background variable, even if it was connected, was the most likely variable to be connected incorrectly, along with *Alpha's report*. Thus, detailed analysis of the Spider errors suggests improving training and help on both logically connected variables and on background variables.

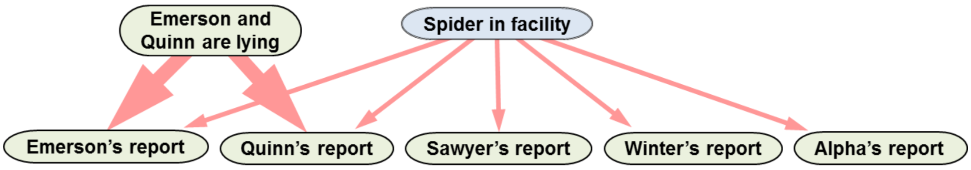


Figure 5: Spider deletion errors (solid arrows) and their relative frequency (via thickness), which include all the gold-standard arrows.

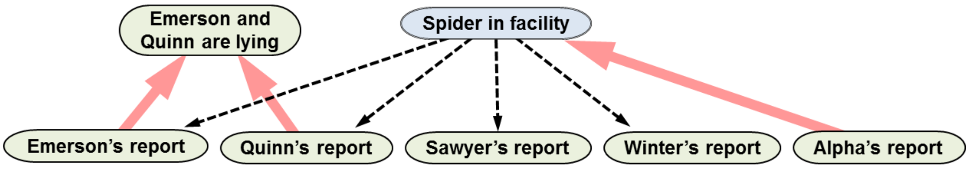


Figure 6: Spider reversal errors (solid arrows) and their relative frequency (via thickness), with other gold-standard arrows (dashed) for context.

7. What specific arrows appeared in the majority of participant models?

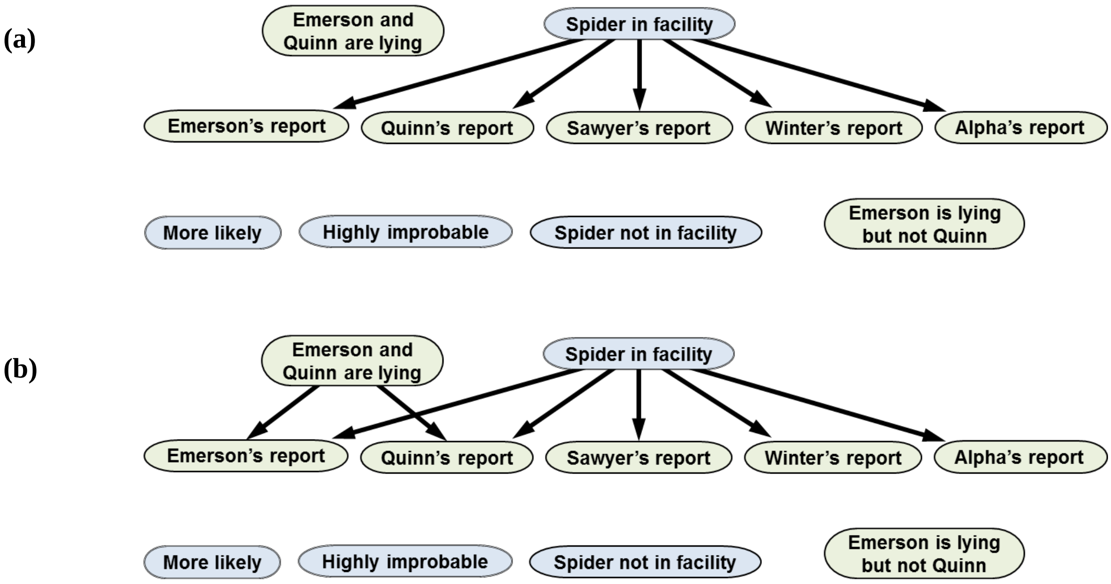


Figure 7: Spider majority-endorsed arrows after (a) Round 1, and (b) Round 2. All arrows were gold-standard and (b) was the complete gold-standard answer.

This is visualised for the Spider problem in Figure 7. In Round 1, five of the seven arrows in the gold-standard model, and no others, were endorsed by the majority of participants. After Round 2, all seven gold-standard arrows were endorsed, and no others. Similar results were achieved in the Black Site problem. Thus, even in Round 1 when very few participants (7–9%) gave the complete gold-standard answer, this criterion would have provided an effective threshold for identifying most of the individual gold-standard arrows. In Round 2, when only a minority of participants (33–42%) gave the complete gold-standard answer, this criterion would have successfully identified the complete gold-standard answer in both problems.

4 Discussion

4.1 BN training and performance

In accordance with BARD's aims, participants received very little upfront BN training and practice by industry standards. It is further validation of this approach that many participants successfully constructed the gold-standard BN structures that were the key to solving these reasoning problems (in this experiment it was the modal response, and in most earlier BARD experiments (Liefgreen, Tešić, Lagnado, 2018; Pilditch, Fenton & Lagnado, 2018; Pilditch, Hahn & Lagnado, 2018; Korb et al., 2019; Pilditch, Fries & Lagnado, 2019) it was also the majority response). Data exploration revealed some general modelling issues where small improvements in training would most likely yield even greater success and consensus on these test problems.

Modelling performance, however, depends on the difficulty of modelling problems. These reasoning problems were difficult enough to defeat participants in other experiments who were unaided by BNs, but the BNs required to solve them are relatively simple (since they involve a small number of binary nodes in a fairly normal causal context). There is, as yet, no standard test for BN modelling difficulty or ability, so we can't quantify more precisely the difficulty of building our problem BNs or the ability achieved through our minimal training.

4.2 Adding other Delphi features

To make the experiment easier and more cost-effective, both the modelling task and the Delphi process were cut down to minimal, central features. Other typical Delphi features could be added that would probably increase the levels of success and consensus. In particular, participants could (i) give a justification for each arrow they included in their model in Round 1; (ii) comment on points of disagreement in Round 2; (iii) have an additional Round 3 to take such comments into account; and (iv) see interim aggregations (in addition to individual responses) so that each participant can focus on their own differences from the prospective group output. These kinds of participant exchanges are encouraged within the BARD process (Nicholson et al., 2019).

4.3 Participant understanding vs heuristics

Although the data shows that participants were able to assess the quality of other models and how to improve their own model in response, we have no further data indicating precisely how these assessments were made. The modelling by individuals in this and other BARD experiments required some good modelling understanding, so we expect that, to a large extent, participants in our group context understood good modelling when they saw it. However, indirect heuristics may also have influenced participants, especially those who were less able to judge model quality directly. Participants may well have favoured solutions that were more popular, or appeared more comprehensive.

In BN modelling, provided that the problem is not too difficult for most participants, the worst solutions are unlikely to be the most popular. Similarly, the worst solutions are unlikely to be the most comprehensive (since good arrows are more likely to be omitted rather than bad arrows added, per our analysis of exploratory question 5). Hence, favouring popular and/or comprehensive solutions is likely to improve the answers of the least able participants, and so (in line with the Theory of Errors) improve the average answer. To the extent that such superficial heuristics contributed to our experimental results, this phenomenon is typical of BARD's Delphi implementation for BN modelling, so our results are a valid test of its performance.

4.4 Extrapolation of results to other BARD Steps

Although our modelling task was limited to specifying model structure, this is central and distinctive to BN construction, and the efficacy of Delphi here was previously untested. Other necessary tasks, such as listing possible relevant factors (variables) or estimating probabilities, are more similar to tasks where Delphi has previously been shown to be effective. Hence, although it awaits direct empirical testing, we expect that Delphi will assist with all BARD Steps.

In our test problems, if participants had been required to develop their own variables from scratch, then it seems likely that less able and/or less industrious participants would have included fewer of the gold-standard variables than they did when these were all suggested and provided by us. However, since our minimal Delphi process achieved a high rate of correction for wrongly omitted arrows, this may well have extrapolated to a high rate of correction for wrongly omitted variables. If participants had been required to enter probabilities, then it's likely there would have been a high degree of consensus – since the required probabilities are explicitly stated in these problems.

4.5 Aggregation methods and automation

Our confirmatory results were about individual performance. However, where (as in BARD, and most Delphi applications) group collaboration is expected to produce a single aggregate result, the performance of the group process depends also on the aggregation procedure. This becomes more critical where there is less consensus, which can occur when the problem is sufficiently challenging (per §4.1), even if a more complete Delphi process is applied (per §4.2). Causal BN structures are complex outputs, and there is no established, best way to 'average' them (as with a single parameter, such as a probability) to produce a representative single structure. Research on the theoretical and applied properties of aggregation rules for BNs has barely begun, and will be a fertile future topic. However, we can make some preliminary comments on the significance of our results here, regarding whether to use (or how to combine) (i) human evaluation vs automation, (ii) prevalence vs rating, and (iii) whole models vs parts.

One BARD configuration is to use a human facilitator, who can choose to use informal, flexible amalgamation methods, e.g. encouraging further discussion between participants on points of disagreement, and hopefully judging when a consensus has formed. Traditional BN elicitation uses a BN expert here, whereas BARD is designed to reduce this human resource overhead by using a facilitator who has no more BN expertise than the analysts. But an alternative BARD configuration is to dispense with a human facilitator altogether, and use only automated methods. This relies on formal amalgamation rules that are clearly and precisely specified.

The whole-model prevalence-based rule (1.1) *accept the majority answer* would not have succeeded amongst our participants, since no whole model commanded a majority after Round 2.[9] This outcome becomes more likely as the complexity of BN answers increases. The weaker rule (1.2) *accept the modal answer* would have successfully chosen the gold-standard answer in both our test problems. However, it is similarly vulnerable: with complex outputs and small samples of responses, there is a

[9] Across our group of participants as a whole, so if real groups had been formed from our participants, on average there would have been no majority of a single model in these groups.

high risk that no single answer will form a high proportion of the responses – indeed, every answer may be different – resulting in either a poorly-supported decision or none at all.

The whole-model rating-based rule (2.1) *accept the highest-rated answer*, or (2.2) *accept the highest-rated answer, provided it is above a minimum quality threshold*, would also have successfully chosen the gold-standard answer in both our test problems. Furthermore, participant ratings of other Round 1 BNs appeared to be as good on average as revised Round 2 BN answers they provided themselves. Advantageously, rating appears to be quicker and easier for participants than revising[10], and even with complex outputs and a small number of participants there will usually be a unique highest-rated answer informed by the sum of all participant opinions. So, this rule for selecting a whole model may well be just as accurate and more reliable than the prevalence-based rules listed above. For the purposes of CREATE assessment, BARD provisionally adopted rule (2.2).

Compared to revising, rating does have a drawback: it only allows participants to choose the best answer available in the previous round, whereas revising allows participants to find an even better answer (if one exists) in the current round. The more complex and difficult the BN is for participants, the more likely it is that initial answers will differ on some points, and that at least one participant will do even better in the next round. However, in such situations it is unlikely that many participants will propose *exactly the same* improved BN in the next round. So, whole-model prevalence-based rules will be unable to immediately select this improved model; it will only be of benefit if there is another round for other participants to endorse it. Hence, the arguments for rating will eventually apply: in earlier rounds, it may be better to ask for revision; but if initiating the final round, then it is better to ask for rating. As usual, the number of rounds should be guided by the likely difficulty of the problem for participants, the diversity of their answers, and the amount of change the latest round has produced.

Part-model approaches, which select components of a final answer, have some advantages over whole-model approaches. Given that BN outputs can be complex with a very large number of possible whole models, simply choosing the best whole model from the small set provided by participants – even after revision in Round 2 – may have relatively low discriminatory power. If good choices can be made

[10] We have no experimental data to confirm this, or estimate how much easier. We do know, however, that both rating and revision require similar comparative assessments of model quality, while rating requires far fewer physical actions.

about parts, and these components assembled into wholes, then this may produce a different and better group model than any of the individual models offered. For example, the part-model prevalence-based amalgamation rule (3.1) *accept every arrow that appears in the majority of responses* would have successfully selected the gold-standard responses to both of our problems after Round 2.

Part-model approaches do have their drawbacks: principally, the risk that the composite result will be inconsistent in some holistic way. Most obviously, models assembled from individually approved arrows could contain cycles. But the low rate of reversal errors we observed suggests that cycles would rarely occur, and in any case, a slightly more complex rule (3.2) could automatically detect and reject cyclical models in favour of similar acyclical alternatives (in a similar way to automated constraint-based causal discovery algorithms, see (Korb & Nicholson, 2011). Nevertheless, the composite model may be inferior in some less obvious way, and it would be more reassuring if there was some human oversight of the final product. An attractive synthesis is for a popular-parts model to be automatically constructed from participant answers (a kind of interim aggregation) and included alongside these individual answers for participant evaluation, at least in the final round. This allows the wisdom of the crowd to be applied to individual components and also applied to reviewing the composite result.

Finally, the lack of correlation between the quality of participants' own answers and of their ratings for others' BNs suggests that both sources of information should also be combined in some way in the amalgamation procedure. Combining all the points above might produce, for example, the rule (4.1) *if answer diversity is high, then conduct another round of revision; but in the final round, select a group model as follows: (a) automatically construct an additional acyclic model based on the prevalence of parts, (b) exclude outlier models with many arrows in unpopular directions to keep the number of candidates manageable, (c) ask all the participants to rate the candidates, and (d) select the highest-rated model provided it has a rating above a minimum threshold*. Plausibly, (a) and (b) would enhance the current BARD amalgamation rule which consists of (c) and (d) only, but empirical testing is clearly required.

5 Conclusions

Causal BNs have become a common AI tool to help humans overcome their well-known difficulties with probabilistic and causal reasoning, and hence make better decisions under uncertainty. However, a major barrier has been the lack of validated methods and software support for groups to build BNs collaboratively, including the amalgamation of diverse opinions into a single group product – particularly if unassisted by a BN expert or human facilitator. Plausibly, the Delphi process can play a central role in the solution, but its effectiveness for various construction steps has been largely untested.

In this context, our novel contributions are:

- We applied a Delphi-style process to amalgamating non-expert opinions about BN structure, in a context where most group members produce high-quality answers but a significant number of errors are made.
- In confirmatory analysis, we found that (i) individuals who viewed the other answers significantly revised and improved the quality of their own answers, and (ii) when these individuals rated the other answers, the answer they rated most highly was significantly better quality than their own initial answer. Since rating is quicker and easier than revising, we suggested that BN amalgamation algorithms may be more efficient if they sometimes use rating (e.g. in the final round) rather than exclusively using revising (as in the traditional Delphi approach).
- Methodologically, (i) we demonstrated the cost-efficient SGRP method for testing a Delphi-style approach, which we can now apply to any of BARD's four BN construction steps, and (ii) we presented novel visualisations for understanding patterns in sets of BN-structure answers.
- In exploratory analysis, we found that (i) improvement was spread widely among participants and across all major error types; (ii) participants who gave worse initial answers were *not* significantly worse at correcting each error (hence such participants made more corrections in total, which reduced the variation in participant scores) and *not* significantly worse at rating the answers of others; (iii) all 'majority' arrows after both Round 1 and Round 2 were part of the gold standard answer, and after Round 2 they provided the complete gold-standard answer (even though the majority of individual participants did not give this answer).

This is a fertile area for further work. The most immediate extensions are to manipulate variables that were held constant in this experiment but, as discussed, might well alter the efficacy of our Delphi process and provide more insight. One is problem difficulty (relative to training), and another related variable is the quality and frequencies of peer answers. In particular, we would like to manipulate the frequency of the gold-standard answer relative to alternatives. Varying gold-standard frequencies can be obtained naturalistically by varying the difficulty of the test problem, but an advantage of SGRP is that we can hold the test problem constant and only vary frequencies in the stimulus set. Extending the Delphi process in the ways described above can also be tested within the SGRP framework and may well prove powerful. Automated amalgamation algorithms can be tested post hoc, as here, or used to generate interim group models as part of an enhanced Delphi process. Finally, we would like to vary the modelling task to measure the efficacy of our Delphi process for other BARD Steps.

References

Arnott, D. (2006). Cognitive biases and decision support systems development: A design science approach. *Information Systems Journal* 16(1): 55–78.

Blascovich, J.; Mendes, W. B.; Hunter, S. B.; & Salomon, K. (1999). Social “facilitation” as challenge and threat. *Journal of Personality and Social Psychology* 77(1): 68–77.

Bolger, F. & Rowe, G. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis* 35(1): 5–11.

Bolger, F.; Rowe, G.; Hamlin, I.; Belton, I.; Crawford, M.; Sissons, A.; Taylor-Browne Lūka, C.; Vasilichi, A. & Wright, G. (2020). The Simulated Group Response Paradigm: A new approach to the study of opinion change in Delphi and other structured group techniques. Submitted to, *Organisational Behavior and Human Decision Processes*; draft at https://tinyurl.com/bard-publications

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence*, pp. 316–334. Cambridge, UK: CUP.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies* 14(1): 75–90.

Buck, R.; Losow, J. I.; Murphy, M. M. & Costanzo, P. (1992). Social facilitation and inhibition of emotional expression and communication. *Journal of Personality and Social Psychology* 63(6): 962–968.

Chee, Y. E.; Wilkinson, L.; Nicholson, A. E.; Quintana-Ascencio, P. F.; Fauth, J. E.; Hall, D.; Ponzio, K. & Rumpff, L. (2016). Modelling spatial and temporal changes with GIS and Spatial and Dynamic Bayesian Networks. *Environmental Modelling and Software* 82: 108–120.

Chidambaram, L. & Tung, L. L. (2005). Is out of sight, out of mind? An empirical study of social loafing in technology-supported groups. *Information Systems Research* 16(2): 149–168.

Corner, A. & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied* 15(3): 199–212.

Dalkey, N. C. (1975). Toward a theory of group estimation. In H.A. Linstone & M. Turoff (Eds.), *The Delphi Method: Techniques and Applications*, pp. 236–261. Reading, MA: Addison-Wesley.

Dodoiu, G.; Leenders, R. T. & van Dijk, H. (2016). A meta-analysis of whether groups make more risky or more cautious decisions than individuals. *Academy of Management Proceedings* 2016(1), 16461.

Etminani, K.; Naghibzadeh, M. & Peña, J. M. (2013). DemocraticOP: A democratic way of aggregating Bayesian network parameters. *International Journal of Approximate Reasoning* 54(5): 602–614.

Fenton, N. & Neil, M. (2000). The Jury Fallacy and the use of Bayesian networks to present probabilistic legal arguments. *Mathematics Today* 36(6): 180–187.

Fenton, N.; Neil, M. & Berger, D. (2016). Bayes and the Law. *Annual Review of Statistics and Its Application* 3: 51–77.

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

Flores, M.; Nicholson, A.; Brunskill, A.; Korb, K. & Mascaro, S. (2011). Incorporating expert knowledge when learning Bayesian network structure: A medical case study. *Artificial Intelligence in Medicine* 53(3): 181–204.

Glass, G. V.; Peckham, P. D. & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research* 42: 237–288.

Hackman, J. R. & Katz, N. (2010). Group behavior and performance. In S. T. Fiske, D. T. Gilbert & G.

Lindzey (Eds.), *Handbook of Social Psychology* (Vol. 2). Hoboken, N.J.: John Wiley & Sons.

Hastie, R. & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review* 112(2): 494–508.

Hahn, U. & Harris, A. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of Learning and Motivation* 61: 41–102.

Hahn, U. & Hornikx, J. (2016). A normative framework for argument quality: Schemes with Bayesian foundation. *Synthese* 193(6):1833–73.

Howson, C. & Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*, 3rd ed. Chicago: Open Court.

Jarvstad, A. & Hahn, U. (2011). Source reliability and the Conjunction Fallacy, *Cognitive Science* 35(4): 682–711.

Johnson, D. D. & Fowler, J. H. (2011). The evolution of overconfidence. *Nature* 477(7364): 317–320.

Kahneman, D.; Slovic, P. & Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. New York: CUP.

Korb, K. (2003). Bayesian informal logic and fallacy. *Informal Logic* 23(2): 21–70.

Korb, K. B.; Hope, L. R. & Nyberg, E. P. (2009). Information-theoretic causal power. In F Emmert-Streib (ed.), *Information Theory and Statistical Learning*, pp 231-65. Springer.

Korb, K. B. & Nyberg, E. P. (2016). Editorial: Analysing arguments using Causal Bayesian Networks. *The Reasoner* 10(4), www.thereasoner.org

Korb, K. B.; Oshni Alvandi, A.; Thakur, S.; Nyberg E. P.; Ozmen, M.; Li, K.; Pearson, R. & Nicholson, A. E. (2019). A collaborative system for Bayesian reasoning: An experimental study. Submitted to *Frontiers in Psychology*; draft at https://tinyurl.com/DP200100040-refs

Korb, K. B. & Nicholson, A. E. (2011). *Bayesian Artificial Intelligence*, 2nd ed. CRC Press, Boca Raton.

Lawlor, D.; Davey Smith, G. & Ebrahim, S. (2004). Commentary: The hormone replacement-coronary heart disease conundrum: Is this the death of observational epidemiology? *International Journal of*

Epidemiology 33(3): 464–467.

Liefgreen, A.; Tešić, M. & Lagnado, D. (2018). Explaining Away: Significance of priors, diagnostic reasoning, and structural complexity. In T. Rogers, M. Rau, X. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pp. 2047–2052. Austin, TX: Cognitive Science Society.

Lichtenstein, S.; Fischhoff, B. & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, pp. 306–334. Cambridge: CUP.

Linstone, H. A. & Turoff, M. (1975). *The Delphi Method: Techniques and Applications*. London: Addison-Wesley.

Morvan, C. & Jenkins, W. J. (2017). *Judgment under Uncertainty: Heuristics and Biases*. London, UK: Macat Library.

Neil, M.; Fenton, N.; Lagnado, D. & Gill, R. (2019). Modelling competing legal arguments using Bayesian model comparison and averaging. *Artificial Intelligence and Law*. 27(4), 403-430

Nicholson, A. E.; Mascaro, S.; Thakur, S.; Korb, K. B. & Ashman, R. (2016). Delphi elicitation for strategic risk assessment. In preparation; available for review upon request.

Nicholson, A. E; Nyberg, E. P.; Korb, K. B.; Mascaro, S.; Thakur, S.; Riley, J.; Wybrow, M.; Morris, S.; Zukerman, I.; Azad, A.; Oshni Alvandi, A.; Bolger, F.; Hahn, U.; Lagnado, D. (2019). BARD: A structured technique for group elicitation of Bayesian networks to support analytic reasoning. In preparation; draft at https://tinyurl.com/bard-publications

Parenté, F. J. & Anderson-Parenté, J. K. (1987). Delphi inquiry systems. In G. Wright & P. Ayton (Eds.), Judgmental Forecasting. Chichester, UK: Wiley.

Pilditch, T. D.; Fenton, N. & Lagnado, D. (2018). The zero-sum fallacy in evidence evaluation. *Psychological Science*, 30(2), 250-260.

Pilditch, T. D.; Fries, A. & Lagnado, D. (2019). Deception in evidential reasoning: Wilful deceit or honest mistake? Accepted to *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. 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.

Rowe, G. & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting* 15(4): 353–375.

Rowe, G.; Wright, G. & Bolger, F. (1991). Delphi: A re-evaluation of research and theory. *Technological Forecasting and Social Change* 39(3): 235–251.

Serwylo, P. (2015). *Intelligently Generating Possible Scenarios for Emergency Management during Mass Gatherings.* Ph.D. Thesis, Monash University.

Sesen, M. B.; Nicholson, A. E.; Banares-Alcantara, R.; Kadir, T. & Brady, M. (2013). Bayesian networks for clinical decision support in lung cancer care. *PLoS One* 8(12), e82349.

Spirtes, P.; Glymour, C. N. & Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press.

Stasser, G. & Vaughan, S. I. (2013). Models of participation during face-to-face unstructured discussion. *Understanding Group Behavior: Consensual Action by Small Groups* 1: 165–192.

Steiner, I. D. (1972). *Group Process and Productivity*. New York, NY: Academic Press.

Stoner, J. A. (1968). Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology* 4(4): 442–459.

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York, NY: Anchor Books.

Turoff, M. (1970). The design of a policy Delphi. *Technological Forecasting and Social Change* 2(2): 149–171.

Villejoubert, G. & Mandel, D. (2002). The inverse fallacy: An account of deviations from Bayes Theorem and the additivity principle. *Memory and Cognition* 30(5): 171–178.

Weber, B. & Hertel, G. (2007). Motivation gains of inferior group members: A meta-analytical review. *Journal of Personality and Social Psychology* 93(6): 973–993.

Wintle, B. & Nicholson, A. (2014). Exploring risk judgments in a trade dispute using Bayesian Networks. *Risk Analysis* 34(6): 1095–1111.

Zajonc, R. B. (1965). Social facilitation. *Science* 149: 269–274.

Zajonc, R. B. (1980). Compresence. In P. B. Paulus (Ed.), *Psychology of Group Influence*, pp. 35–60. Hillsdale, NJ: Erlbaum.

Supplementary Materials

Details of the stimulus materials can be found here:

https://tinyurl.com/StructureDelphi-Supplements