Explaining away: significance of priors, diagnostic reasoning, and structural complexity

Alice Liefgreen[1,*] (alice.liefgreen.15@ucl.ac.uk) **& Marko Tešić**[2,*] (mtesic02@mail.bbk.ac.uk)
David Lagnado[1] (d.lagnado@ucl.ac.uk)

[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
* equal contribution

Abstract

Recent research suggests that people do not perform well on some of the most crucial components of causal reasoning: probabilistic independence, diagnostic reasoning, and explaining away. Despite this, it remains unclear what contexts would affect people's reasoning in these domains. In the present study we investigated the influence of manipulating priors of causes and structural complexity of Causal Bayesian Networks (CBNs) on the above components. Overall we found that participants largely accepted the priors and understood probabilistic independence, but engaged in inaccurate diagnostic reasoning and insufficient explaining away behavior. Moreover, the effect of manipulating priors on participants' performance in diagnostic reasoning and explaining away was significantly larger in a structurally less complex CBN than in a structurally more complex CBN.

Keywords: Explaining Away; Diagnostic Reasoning; Prior probability; Causal Bayesian Networks; Network Complexity; Interpretations of Probability; Propensity

Introduction

Explaining away is a pattern of inference that occurs in situations where independent causes compete to account for an effect and is best understood in two stages.[1] First, an effect is observed (e.g. a barn is burned down) and via diagnostic reasoning we update (increase) the probability of each cause (e.g. careless smoking and faulty electrical wiring). Subsequently, once we observe the presence of one cause (faulty electrical wiring) we update (decrease) the probability of the other cause (careless smoking), and we say that faulty electrical wiring explains away the the observation of the burnt barn. Conversely, if we were to find out about the absence of one cause (good electrical wiring), we would update (further increase) the probability of the other cause (careless smoking). This pattern of inference is found in a wide range of contexts including social attribution, medical diagnosis, and legal domains (Kelley, 1973; Rottman & Hastie, 2016), rendering it a pivotal building block of causal reasoning.

Situations that involve explaining away can be modeled using Causal Bayesian Networks (CBNs) (Pearl, 2009). Figure 1a illustrates a common effect model, where two causes (modeled as variables C_1 and C_2) independently cause an effect (modeled as a variable E). To fully parametrize the CBN one needs to specify the prior probabilities of the two causes (i.e. $P(C_1 = 1)$ and $P(C_2 = 1)$[2]) as well as the conditional probabilities of the effect given the presence and/or absence of each cause (i.e. $P(E = 1 \mid C_1 = 1, C_2 = 1)$, $P(E = 1 \mid C_1 = 1, C_2 = 0)$, $P(E = 1 \mid C_1 = 0, C_2 = 1)$, $P(E = 1 \mid C_1 = 0, C_2 = 0)$). Following this, the normative theory of CBNs dictates how one should compute and infer the probability of any variable(s) in the network being present/absent given any other variable(s) being present/absent.

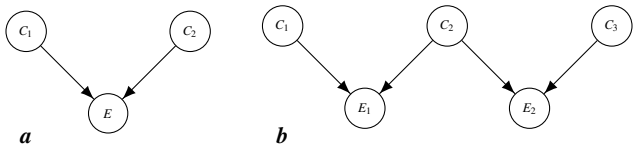

Figure 1: (a) 3-node CBN and (b) 5-node CBN

Note that the common effect model does not always lead to explaining away and that the explaining away pattern emerges only when the following inequality applies (Wellman & Henrion, 1993):

$$\frac{P(E = 1 \mid C_i = 0, C_j = 0)}{P(E = 1 \mid C_i = 0, C_j = 1)} < \frac{P(E = 1 \mid C_i = 1, C_j = 0)}{P(E = 1 \mid C_i = 1, C_j = 1)} \quad (1)$$

From Inequality (1) it follows (see Morris & Larrick, 1995):

$$\begin{aligned} P(C_i = 1 \mid E = 1, C_j = 1) &< P(C_i = 1 \mid E = 1) \\ &< P(C_i = 1 \mid E = 1, C_j = 0) \end{aligned} \quad (2)$$

The inequalities in (2) follow the general intuition of explaining away reflected in the aforementioned barn example and serve as a definition of explaining away (see Rehder & Waldmann, 2017; Rottman & Hastie, 2016).

Studies investigating people's ability to correctly engage in explaining away, applying the 'stricter' definition presented above by the inequalities in (2), have so far yielded mixed findings. Overall, results suggest that people tend to explain away either 'insufficiently' or not at all (Fernbach & Rehder, 2013; Morris & Larrick, 1995; Rehder & Waldmann, 2017; Rottman & Hastie, 2016). Some studies have also reported results opposite to explaining away, i.e. $P(C_i = 1 \mid E = 1, C_j = 1) > P(C_i = 1 \mid E = 1, C_j = 0)$ (Fernbach & Rehder, 2013; Rehder, 2014). Hence, it remains unclear under what circumstances explaining away displays itself.

[1]A related concept to explaining away is discounting (for difference between the concepts see Rehder & Waldmann, 2017; Rottman & Hastie, 2014). In this paper, however, we are only going to address explaining away.

[2]$C_i = 1$ and $E = 1$ indicate that the cause C_i and effect E are present, respectively.

One of the major shortcomings in the extant literature is that very few studies allow direct comparisons to be made between participants' inferences and the normative model since the prior probabilities were not explicitly established. Even among the studies where priors are given, participants were expected to infer or to calculate them (Rehder & Waldmann, 2017; Rottman & Hastie, 2016). Moreover, it is unclear whether in these studies participants accepted the priors given to them. Following Rottman and Hastie's (2016) approach, we manipulated the prior probabilities of the causes so that they were either 'medium' or 'low'. We addressed the aforementioned shortcomings by explicitly giving prior probabilities to participants without the need for calculations and by asking direct questions to gauge whether these priors were accepted.

Studies that report participants' insufficient or inexistent explaining away, likewise report violations of the Markov parental condition, i.e. $P(C_i = 1 \mid C_j = 1) > P(C_i = 1 \mid C_j = 0)$ (Rehder, 2014; Rehder & Waldmann, 2017; Rottman & Hastie, 2016). In common effect structures such as that of Figure 1a, the Markov parental condition stipulates probabilistic independence between the two causes (C_1 and C_2) when the state of the effect (E) is unknown (Pearl, 2009). Studies that draw conclusions on explaining away whilst reporting a violation of probabilistic independence are problematic, since this independence is one of the necessary conditions for explaining away. It has been proposed that the violation could be due to participants assuming that variables are related by another implicit underlying common cause, or holding a priori beliefs about the correlation of the causes (Rehder & Burnett, 2005). In the present experiment we reverted to a simple coin-tossing scenario (Pearl, 2009) that should minimize participants' susceptibility to these issues as coin toss outcomes are intuitively independent of one another.

Another potential drawback apparent in the existing explaining away literature relates to the manner with which probability estimates are elicited from participants. For instance, studies including questions relating to diagnostic reasoning often request probability estimates of a cause given an effect $P(C_i \mid E)$, but do not ask participants to consider the relation and direction of change of this probability compared to the prior probability of the cause. To rectify this issue, in the present study, participants were asked a *qualitative* question, e.g. whether $P(C_1 \mid E)$ is less than, greater than or equal to $P(C_1)$, followed by a *quantitative* question asking them to provide a probability estimate of e.g. $P(C_1 \mid E)$. Given that explaining away is a relational concept (inequalities in (2)), asking both types of questions (i) ensures that we remained true to the strict definition of explaining away, (ii) facilitated the elicitation of probabilistic judgments in the context of explaining away, and (iii) increased the informativeness of our data.

We also explored intercausal reasoning in a more complex 5-node structure (see Figure 1b), originally introduced by Wellman and Henrion (1993). This allowed us to determine whether the existing findings of explaining away are restricted to 3-node structures (Figure 1a), or if they also extend to more complex structures. Given the added complexity, we expected the performance of participants who are asked to reason with the 5-node structure to explain away less than participants reasoning with the 3-node structure. Introducing a 5-node structure (Figure 1b) additionally enabled us to distinguish between *direct* explaining away—i.e. explaining away discussed thus far— and *chained* explaining away (Wellman & Henrion, 1993), where, assuming that we know $E_1 = 1$ and $E_2 = 1$, learning $C_1 = 1$ decreases the probability of $C_2 = 1$, which in turn increases the probability of $C_3 = 1$.

Experiment Overview

In the present experiment we investigated the influence of manipulating priors and structural complexity on independence, diagnostic reasoning, and explaining away. Participants were required to reason with one of four models.[3] Model 1 was a 3-node structure where $P_1(C_1 = 1) = P_1(C_2 = 1) = 0.5$, Model 2 was a 3-node structure where $P_1(C_1 = 1) = 0.2$ and $P_1(C_2 = 1) = 0.1$, Model 3 was a 5-node structure where $P_3(C_1 = 1) = P_3(C_2 = 1) = P_3(C_3 = 1) = 0.5$ and finally Model 4 was a 5-node structure where $P_4(C_1 = 1) = 0.1$, $P_4(C_3 = 1) = 0.2$, and $P_4(C_3 = 1) = 0.1$. In all four models the presence of one cause entailed the presence of the effect—$P(E = 1 \mid C_i = 1, C_j = 1) = P(E = 1 \mid C_i = 1, C_j = 0) = P(E = 1 \mid C_i = 0, C_j = 1) = 1$—and absence of both causes entailed absence of the effect—$P(E = 1 \mid C_i = 0, C_j = 0) = 0$. This resulted in the base rate for the effect $P(E = 1) = 0.75$ in models with medium priors and $P(E = 1) = 0.28$ in models with low priors. We predicted that these deterministic relations between causes and effect(s) would further facilitate both diagnostic and explaining away reasoning.

Methods

Participants

A total of 204 participants ($N_{\text{MALE}} = 81$, $M_{\text{AGE}} = 37$ years) were recruited from Prolific Academic (`www.prolific.ac`). All participants were native English speakers who gave informed consent and were paid $1.25 for partaking in the present study, which took on average 16.7 minutes to complete. Seven participants were excluded as they answered incorrectly to the catch trial, leaving a total of 197 participants in the analyses.

Design and Materials

A mixed-subjects design was adopted. Participants were randomly assigned to one of four groups ($N_{\text{Group 1}} = 51$, $N_{\text{Group 2}} = 47$, $N_{\text{Group 3}} = 49$, $N_{\text{Group 4}} = 50$). They were all given the same coin-tossing cover story wherein simultaneously tossed coins (C; binary variable, assumes the value of either H or T) in separate rooms could lead to a light bulb (LB) switching on in a different unit depending on the outcome of the toss (if at least one coin lands Heads, the

[3] By a model we mean a structure that has been fully parameterized.

light bulb turns on). Each Group i was assigned a Model i (where $i \in \{1,2,3,4\}$) and asked to complete an inference questionnaire (Models 1 and 2, $N_{\text{questions}} = 10$; Models 3 and 4, $N_{\text{questions}} = 17$).

Procedure

Participants in each group were initially presented with the cover story and subsequently given information on their model (i.e. variables present, the priors of coins, and causal relationships within the model). This was done in both textual form and in visual form (graphical representation). Moreover, participants were provided with a textual account by which each cause could independently bring about the effect.

Subsequently, each participant proceeded to complete the inference questionnaire. Although the questions varied among groups, they were all nested within the same inference types (see Table 1) and were presented in the same order. Participants firstly answered questions regarding prior probabilities of causes, secondly regarding independence of causes, thirdly regarding diagnostic reasoning, and finally regarding explaining away. Inferences besides those present in the table were asked, but will not be discussed at present.

To investigate participants' diagnostic and explaining away reasoning, we asked questions in both qualitative and quantitative formats. For example, participants in Groups 1 and 2, after finding out that the light bulb is on, were asked both a qualitative diagnostic reasoning question: "Does the probability that **Coin 1** landed Heads change after you find out that the light bulb turned on?" and a quantitative diagnostic reasoning one: "What do you now think is the probability that **Coin 1** landed Heads?". Qualitative questions were presented in a multiple choice format with three options: the probability increases, decreases, and stays the same. Quantitative questions required participants to provide their probabilistic estimate on a slider with a scale ranging from 0% to 100%. Questions relating to prior probabilities of the causes were asked only in the quantitative format, whereas questions relating to independence of the causes were asked solely in the qualitative format. In addition, some questions prompted participants to provide written explanations for their answers.

Throughout the questionnaire, evidence about the state of the variables was provided to participants both textually (e.g. "You walk into Room 1 and see that the light bulb is on.") as well as with an updated graphical representation of the model.

Results

Overall Performance

An overall performance score was computed per participant[4] and subsequently converted to a percentage score. An average percentage score was obtained for each group. A Kruskal-Wallis H test showed no statistically significant difference in the average percentage score across the four groups ($M_{\text{Group 1}} = 33.2\%$, $M_{\text{Group 2}} = 44.4\%$, $M_{\text{Group 3}} = 34.4\%$, $M_{\text{Group 4}} = 33.7\%$, $\chi^2(3) = 3.37$, $p = 0.34$).

To determine whether the average percentage scores differed from the chance level we ran 10,000 simulations per model.[5] These allowed us to conclude that, within each model, obtaining an average percentage score greater than or equal to that of participants' was highly unlikely ($p < 0.0001$). According to the simulations, the average percentage score at the chance level was 23.8% for Models 1 and 2 and 24.3% for Models 3 and 4.

Priors

Fisher's Exact test showed a significant difference in the percentage of participants who correctly answered[6] all questions related to priors across the four models, $p < 0.0001$. More specifically, pairwise comparisons[7] showed that Group 1 (100%) significantly differed from both Group 2 (47%), $p < 0.0001$ and Group 4 (52%), $p < 0.0001$. In addition, Group 3 (94%) significantly differed from both Group 2, $p < 0.0001$ and Group 4, $p < 0.0001$. This suggests that participants accepted priors of causes given to them notably more when they were medium, than when they were low. However, even in groups that were given low priors, a notable portion of participants simply confused the probability of Tails with the probability of Heads (which can be seen from Figure 2).

Independence

The percentage of participants who correctly answered all questions related to independence (see Table 1) was 96% in Group 1, 81% in Group 2, 100% in Group 3, and 94% in Group 4. These were notably higher than the chance level of $(1/3)^2 \approx 11\%$ (i.e. randomly selecting the correct answer to each of the two multiple choice questions with three options) in Groups 1 and 2 and $(1/3)^3 \approx 4\%$ (i.e. randomly selecting the correct answer to each of the three multiple choice questions with three options) in Groups 3 and 4. Overall, the vast majority of participants correctly reported probabilistic independence between causes regardless of the model they were assigned to. This implies that there was no violation of the Markov condition in any of the groups.

Diagnostic Reasoning

The percentage of participants who correctly answered both *qualitative* questions related to diagnostic reasoning (see Table 1) was 6% in Group 1, 42.5% in Group 2, 20.4% in Group

[4]For questions about prior probabilities, participants received 0.5 (0.25) points if their answer was within $\pm 5\%$ ($\pm 10\%$) of the stated prior probability; otherwise, they received 0 points. For questions about independence, participants received 1 point if correctly answered, otherwise 0. For all other *qualitative* questions, participants received 2 points if correctly answered, otherwise 0. To avoid artificial inflation of scores, whereby a participant gave a (close to) correct probability estimate but for the wrong reason, *quantitative* questions were conceived as bonus point questions: a participant received 1 (0.5) point if his/her answer was within $\pm 5\%$ ($\pm 10\%$) of the normative answer *and* if s/he has correctly answered the corresponding qualitative question; otherwise, s/he received 0 points.

[5]One simulation per Model i is $N_{\text{Group } i}$ agents (number of participants in Group i) randomly choosing answers, which were then scored and the average percentage score for the whole Model i is taken.

[6]An answer is considered to be correct if it falls within $\pm 5\%$ interval of the stated prior probability.

[7]All pairwise comparisons used adjusted $\alpha = 0.008$.

Table 1: Key inferences per group.

Inference Type	Key Inferences	
	Group 1 and 2	**Group 3 and 4**
Priors	$P_{1,2}(C_1 = \mathrm{H})$, $P_{1,2}(C_2 = \mathrm{T})$	$P_{3,4}(C_1 = \mathrm{H})$, $P_{3,4}(C_2 = \mathrm{T})$, $P_{3,4}(C_3 = \mathrm{H})$
Independence	$P_{1,2}(C_2 = \mathrm{H} \mid C_1 = \mathrm{H})$, $P_{1,2}(C_1 = \mathrm{H} \mid C_2 = \mathrm{T})$	$P_{3,4}(C_2 = \mathrm{H} \mid C_1 = \mathrm{H})$, $P_{3,4}(C_3 = \mathrm{H} \mid C_2 = \mathrm{T})$, $P_{3,4}(C_3 = \mathrm{H} \mid C_2 = \mathrm{H})$
Diagnostic Reasoning	$P_{1,2}(C_1 = \mathrm{H} \mid LB = \mathrm{on})$, $P_{1,2}(C_2 = \mathrm{H} \mid LB = \mathrm{on})$	$P_{3,4}(C_1 = \mathrm{H} \mid LB_1 = \mathrm{on})$, $P_{3,4}(C_2 = \mathrm{H} \mid LB_1 = \mathrm{on})$
Explaining Away	$P_{1,2}(C_1 = \mathrm{H} \mid LB = \mathrm{on}, C_2 = \mathrm{H})$ $P_{1,2}(C_2 = \mathrm{H} \mid LB = \mathrm{on}, C_1 = \mathrm{H})$ $P_{1,2}(C_2 = \mathrm{H} \mid LB = \mathrm{on}, C_1 = \mathrm{T})$	**Direct:** $P_{3,4}(C_2 = \mathrm{H} \mid LB_1 = \mathrm{on}, C_1 = \mathrm{H})$ $P_{3,4}(C_2 = \mathrm{H} \mid LB_1 = \mathrm{on}, LB_2 = \mathrm{on}, C_1 = \mathrm{H})$ $P_{3,4}(C_2 = \mathrm{H} \mid LB_1 = \mathrm{on}, LB_2 = \mathrm{on}, C_1 = \mathrm{T})$ **Chained:** $P_{3,4}(C_1 = \mathrm{H} \mid LB_1 = \mathrm{on}, LB_2 = \mathrm{on})$ $P_{3,4}(C_3 = \mathrm{H} \mid LB_1 = \mathrm{on}, LB_2 = \mathrm{on}, C_1 = \mathrm{H})$

3, and 30% in Group 4 (the chance level is $(1/3)^2 \approx 11\%$). Fisher's Exact test showed that these percentages significantly differed across the four groups, $p = 0.0001$. Pairwise comparisons showed the only significant differences ($\alpha = 0.008$) were between Group 1 and both Group 2, $p < 0.0001$ and Group 4, $p = 0.002$. The percentage of participants who correctly[8] answered both *quantitative* questions related to diagnostic reasoning was 0% in Group 1, 11% in Group 2, 10% in Group 3 and 8% in Group 4. Fisher's Exact test showed no significant difference between the groups, $p = 0.07$.

These findings suggest that manipulating prior probabilities affected participants' accuracy on qualitative questions relating to diagnostic reasoning more in the simple 3-node model than in the complex 5-node model. The same effect was not found in relation to performance on quantitative diagnostic reasoning questions.

Explaining Away

Direct The percentage of participants who correctly answered all three *qualitative* questions related to direct explaining away (see Table 1) was 2% in Group 1, 25.5% in Group 2, 6% in Group 3, and 4% in Group 4 (the chance level is $(1/3)^3 \approx 4\%$). Fisher's Exact test showed that these percentages significantly differed across the four models, $p = 0.0004$. Pairwise comparisons revealed significant differences ($\alpha = 0.008$) between Group 2 and both Group 1, $p = 0.006$ and Group 4, $p = 0.003$.

Hence, similarly to the findings regarding diagnostic reasoning, there was a significantly larger difference in accuracy on *qualitative* questions relating to direct explaining away between groups reasoning with 3-node models than between groups reasoning with 5-node models.

A repeated measures ANOVA with a Greenhouse Geisser correction showed a significant difference in the average probability estimates related to *quantitative* direct explaining away questions within Group 1, $F(1.5, 75.6) = 22.7$, $p < 0.0001$; within Group 2, $F(2.3, 105.6) = 31.56$, $p < 0.0001$; within Group 3, $F(2.05, 98.5) = 63.3$, $p < 0.0001$ and within Group 4, $F(2.2, 109.9) = 17.4$, $p < 0.0001$. Post-hoc paired t-tests on pairs of inferences (see Figure 3) allowed us to obtain 95% CI of the difference in the average probability estimates between inferences[9] (see Table 2).

Table 2: Within group explaining away.

Norm. diff. = normative difference, Emp. diff. = empirical difference, 95% CI of emp. diff. = 95% CI of empirical difference.

Inferences[10]	Norm. diff.	Emp. diff.	95% CI of emp. diff.
Group 1			
A − B	17	−1.2	[−5, 2.4]
C − B	50	24.2	[14.1, 34.3]
A′ − B′	17	0.8	[−3.3, 4.9]
Group 2			
A − B	26	11.3	[4.2, 18.4]
C − B	90	46.2	[32.4, 59.9]
A′ − B′	51	13.9	[7.3, 20.5]
Group 3			
D − E	17	5.7	[2.8, 8.5]
F − E	50	33.3	[26.6, 39.8]
D′ − E′	13	2.1	[−0.7, 4.9]
Group 4			
D − E	51	6.3	[−0.1, 12.7]
F − E	80	31.9	[19.4, 44.4]
D′ − E′	25	7.3	[3.2, 11.3]

Comparing answers to *quantitative* questions showed there was no sufficient explaining away in any group, since the normative difference was not included in any of the 95% CI in Table 2. However, different levels of insufficiency were found between the groups. Only in Group 2 was the amount of explaining away greater than zero in all three comparisons.

Welch's ANOVA was run on the average estimate given to inferences C/F (see Figure 3) across all models ($M_{\text{Group 1}} = 76.9\%$, $M_{\text{Group 2}} = 78.3\%$, $M_{\text{Group 3}} = 82.9\%$, $M_{\text{Group 4}} = 67.1\%$). Results showed no significant difference in participants' average probability estimate on these inferences between models, $F(3, 105.7) = 2.3$, $p = 0.08$. In addition, each group's average estimate was significantly lower than the re-

[8] Answer falls within ±5% interval of the normative answer.

[9] CI interpretation: lower bound = whether the amount of explaining away is significantly higher than zero; upper bound = whether the amount of explaining away is significantly lower than the normative amount (see Rottman and Hastie, 2016).

[10] $\mathrm{A}' : P_{1,2}(C_1 = \mathrm{H} \mid LB = \mathrm{on})$, $\mathrm{B}' : P_{1,2}(C_1 = \mathrm{H} \mid LB = \mathrm{on}, C_2 = \mathrm{H})$, $\mathrm{D}' : P_{3,4}(C_2 = \mathrm{H} \mid LB_1 = \mathrm{on}, LB_2 = \mathrm{on})$, $\mathrm{E}' : P_{3,4}(C_2 = \mathrm{H} \mid LB_1 = \mathrm{on}, LB_2 = \mathrm{on}, C_1 = \mathrm{H})$. A', B', D', and E' are not illustrated in Figure 3.

spective normative answer to inferences C/F, namely 100% (see Figure 3).

Chained The percentage of participants who correctly answered all *qualitative* questions related to chained explaining away (see Table 1) was 4% in Group 3 and 8% in Group 4 (the chance level is $(1/3)^2 \approx 11\%$). Fisher's Exact test showed that these percentages did not significantly differ, $p = 0.36$. This suggests that manipulating priors did not affect participants' accuracy in qualitative questions relating to chained explaining away. In addition, after collapsing Groups 3 and 4, an exact McNemar's test showed no significant difference between the proportion of participants who correctly answered qualitative questions related to *direct* explaining away (5%) and those who correctly answered qualitative questions related to *chained* explaining away (4%), $p = 1$.

Finally, within each group, we identified participants who correctly answered both *qualitative* questions related to diagnostic reasoning and compared their average performance score on subsequent explaining away questions, to that of participants who incorrectly answered at least one *qualitative* question related to diagnostic reasoning. A Welch t-test showed a significant difference between the average performance scores within Group 2, $t(29) = 4.76$, $p < 0.0001$; Group 3, $t(10) = 3.5$, $p = 0.006$ and Group 4, $t(22) = 3.4$, $p = 0.002$. No significant difference was found within Group 1, $t(2) = 1.52$, $p = 0.26$. This suggests that correct *qualitative* diagnostic reasoning is predictive of better performance on explaining away, although this does not hold for participants who reasoned with a simple 3-node structure with medium priors (Group 1), which can be attributed to the fact that only 3 participants in Group 1 correctly answered both diagnostic reasoning questions.

Discussion

In the present study we investigated the impact of manipulating prior probabilities of the causes and structural complexity on independence, diagnostic reasoning, and explaining away.

Overall, participants accepted priors of causes in all conditions. In stark contrast to the existing literature (Rehder, 2014; Rehder & Waldmann, 2017; Rottman & Hastie, 2016), we found no violation of the Markov condition in any of the groups, implying that one of the crucial assumptions of explaining away was satisfied across all groups. This suggests that an understanding of probabilistic independence might be contingent on the particular cover story used (i.e. coin tossing) and/or how participants were asked about the independence (i.e. qualitative relational questions).

Manipulating prior probabilities significantly affected performance on *qualitative* questions related to diagnostic reasoning as well as direct explaining away, in 3-node models but not in 5-node models. More specifically, Group 2 performed better than all other groups in questions related to *qualitative* diagnostic reasoning and explaining away. However, more pronounced explaining away behavior was expected in Group 2 since Model 2 normatively required the largest amount of explaining away (see Figure 3).

In line with the previous research on explaining away (Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016), insufficient direct explaining away was observed within all groups when comparing answers to *quantitative* questions. Notably, participants reasoning with a 3-node model with low priors (Group 2) performed better than participants in other conditions in *quantitative* explaining away questions (see Table 2), mirroring the findings of Rottman and Hastie (2016). Taken together, participants' answers to both qualitative and quantitative questions seem to suggests that a 3-node structure with low priors bolsters accurate explaining away reasoning.

By adding structural complexity we were able to investigate the unexplored phenomenon of chained explaining away (Wellman & Henrion, 1993). Amongst participant groups reasoning with 5-node models, we did not find a difference in performance between direct and chained explaining away.

Another interesting finding relates to the 'diagnosticity' of diagnostic reasoning. Namely, our results are the first to suggest that correct *qualitative* diagnostic reasoning is predictive of better performance on explaining away. Additionally, we found that a significant proportion of participants in each group remained at their initial (prior) probability estimates of the causes throughout the questionnaire, i.e. did not update the probabilities of the causes given evidence. This was most pronounced in Group 1 where 76% of participants did not update the probabilities of the causes; in the other groups, although less pronounced, the proportion was still surprisingly high: 38% in Group 2, 49% in Group 3, and 32% in Group 4.

A possible explanation of the findings discussed thus far is that a significant number of participants interpreted the probabilities in our cover story as propensities. A propensity can be thought of as a tendency of a physical system to produce a certain outcome (see Hájek, 2012; Popper, 1959). Since our cover story included a coin-tossing mechanism that tosses coins with an established probability for Heads/Tails, it is plausible that some participants interpreted updating probabilities in diagnostic reasoning and explaining away questions as a request to update the coin propensities that were first given to them (i.e. to update the tendency of a coin-tossing mechanism to produce a certain outcome). As intuitively these coin biases (propensities) stay the same, there is no incentive to change these—that people intuitively differentiate amongst the interpretations of probability has been suggested in Kahneman and Tversky (1982) and evidentially supported in Fox and Ülkümen (2011). Moreover, the propensity interpretation may be even more pronounced given the phrasing of the questions (see Procedure) in our questionnaire (see Ülkümen, Fox, & Malle, 2016). This propensity interpretation would then account for the aforementioned significant portion of participants who did not update the probabilities throughout the questionnaire.

Further evidence for the hypothesis that a significant number of participants interpreted probabilities as propensities,

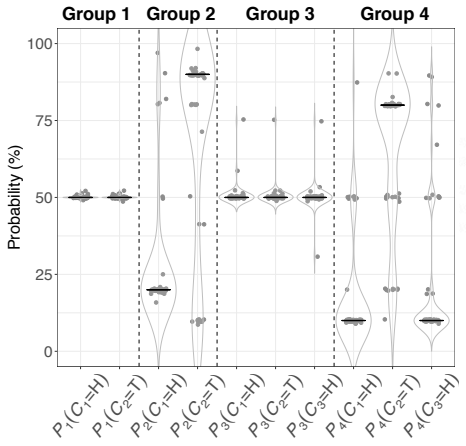

Figure 2: Distribution of prior probabilities. Black lines are stated priors.

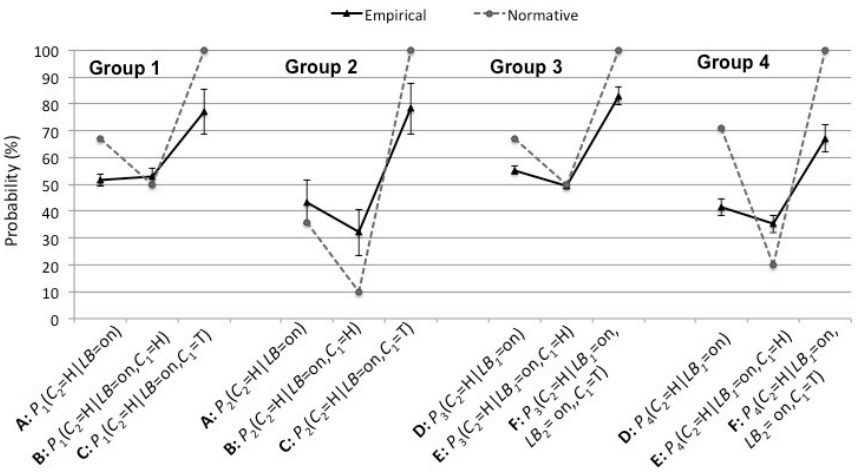

Figure 3: Explaining away inferences. Error bars are 95% confidence intervals.

is performance on explaining away inferences C/F (see Figure 3). Here, a significant number of participants did not update the probability to 100% despite the fact that it was a logical and intuitive update to make. In addition, this hypothesis could also partly explain overall poor performance on diagnostic reasoning and explaining away across groups. It would then follow that the smaller portion of participants who exhibited accurate diagnostic reasoning and explaining away, did not interpret probabilities as propensities, but rather might have adopted an epistemic (subjective) interpretation. Results from Study 1 found in Morris and Larrick (1995) where participants interpreted probabilities in an epistemic (subjective) way and did fairly well on diagnostic reasoning and explaining away questions support this claim.

However, we believe that participants' differential interpretation of probability is only part of the explanation of poor/good performance in diagnostic reasoning questions in the present study. Namely, even participants who were given low priors and who correctly answered qualitative diagnostic reasoning questions by indicating an increase in probability, gave quantitative estimates of 50% for all causes or a more sophisticated 66% and 33%, depending on the priors of the causes. This then suggests that if the prior probabilities of the causes were high (e.g. 90% and 80%), following the same reasoning, participants would decrease these probabilities, contrary to the normative model. This hypothesis should be tested in further research. In addition, the above discussion points can inform a novel study whereby participants adopt a solely epistemic interpretation of probability when making inferences on independence, diagnostic reasoning, and explaining away.

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/phbns/`.

References

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, *4*(1), 64–88.

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Kirkebøen, & H. Montgomery (Eds.), *Essays in judgment and decision making* (pp. 21–35). Oslo, Norway: Universitetsforlaget.

Hájek, A. (2012). Interpretations of probability. In *The stanford encyclopedia of philosophy.*

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143–157.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107–128.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331-355.

Pearl, J. (2009). *Causality*. Cambridge university press.

Popper, K. R. (1959). The propensity interpretation of probability. *The British journal for the philosophy of science*, *10*(37), 25–42.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, *72*, 54–107.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive psychology*, *50*(3), 264–314.

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, *45*(2), 245–260.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological bulletin*, *140*(1), 109–139.

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive psychology*, *87*, 88–134.

Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of experimental psychology: General*, *145*(10), 1280–1297.

Wellman, M. P., & Henrion, M. (1993). Explaining "explaining away". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(3), 287–292.