Explaining away and the propensity interpretation of probability: the case of unequal priors

ALICE LIEFGREEN
Department of Experimental Psychology,
University College London
alice.liefgreen.15@ucl.ac.uk

MARKO TEŠIĆ
Department of Psychological Sciences,
Birkbeck University
mtesic02@mail.bbk.ac.uk

Explaining away is a pattern of inference that occurs in situations where independent causes compete to account for an effect. Empirical studies have found that people 'insufficiently' explain away. In this paper we explore whether this insufficiency could be partly due to people's different interpretations of probabilities. In particular, we tested people on reasoning tasks involving unequal priors of causes and we provide evidence indicating that some people may interpret probabilities as propensities, which would then drive the insufficiency effect of explaining away.

KEYWORDS: Causal Bayesian networks; Causal inference; Diagnostic reasoning; Evidential reasoning; Explaining away; Probability interpretation; Propensity

1. INTRODUCTION

Judgments and inferences that are reliant on beliefs about how events or items of information are causally related to each other are extremely ubiquitous in people's daily and professional lives. The vast majority of these causal judgments occur *under uncertainty*. Consider for example a scenario in which a social worker is trying to ascertain whether action should be taken to remove a child displaying bruises from the custody of his parents under the suspicion that he is being physically abused. The social worker, however, knows from her experience that the bruises could also be the product of a blood disorder termed 'haemophilia'. Upon

observing the bruises, she should then *increase* the probability of each potential cause as bruises are indicative of each one. After a medical test, the social worker learns that the child definitely suffers from haemophilia. Given this new piece of information, she should now *decrease* the probability of the child being physically abused, since haemophilia is sufficient to explain the bruises. If, however, the medical test had revealed that the child definitely *did not* suffer from haemophilia, then the probability of him being physically abused would *further increase* as a result. This scenario illustrates a pervasive pattern of reasoning known as 'explaining away'. In more general terms, explaining away occurs in situations in which multiple independent causes (e.g. physical abuse and haemophilia) compete to explain a common effect (e.g. bruises). After observing the occurrence of the effect, the probability of the two causes increases (step 1). Subsequently, after learning of the occurrence of one the probability of the alternative cause(s) decreases (step 2a). If, conversely, we learned that a cause *did not* happen, the probability of the other cause(s) further increases (step 2b).

1.1. Explaining away: Normative account

Over the past few decades, patterns of inference in causal reasoning such as explaining away have been modelled in the cognitive sciences utilising graphical models called 'Causal Bayesian Networks' (CBNs). These can be used to represent probabilistic knowledge in a graphical manner (for overview see Pearl, 2009; Neapolitan, 2003).

The computational machinery of CBNs, grounded in probability theory, allows one to perform exact quantitative computations of the probability of any random variable(s) in the network being present/absent given the presence/absence of any other variables.

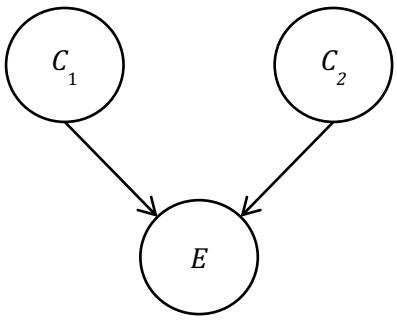

Figure 1: A CBN model of explaining away

Consider the graph in Figure 1, consisting of three nodes representing three random variables: two causes, C_1 and C_2, and one common effect, E[1]. Situations involving explaining away can be modelled utilising common-effect CBNs such as the one in Figure 1 (see Pearl, 1998; 2009). For example, we could model the aforementioned example by representing physical abuse as C_1, haemophilia as C_2 and finally the bruises on the body as E. The two causes are (unconditionally) independent when we do not know whether the child has bruises on his body or not, which follows our intuitions that physical abuse and haemophilia cannot probabilistically influence each other, *before* learning anything about the bruises. However, this is dependent on the network parameterization.

In order for common-effect CBNs to lead to the pattern of explaining away described earlier, they need to be parameterized such that the following inequality holds (see Wellman, 1993):

$$\frac{P\left(E\middle|\neg C_i, \neg C_j\right)}{P\left(E\middle|\neg C_i, C_j\right)} < \frac{P(E|C_i, \neg C_j)}{P\left(E\middle|C_i, C_j\right)} \tag{1}$$

for $i, j \in \{1, 2\}$. From Inequality (1) it follows (see Griffiths, 2001; Morris & Larrick, 1995):

$$P(C_i|E, C_j) \; < \; P(C_i|E) < \; P(C_i|E, \neg C_j) \tag{2}$$

The inequalities in (2) comply with the general intuition of explaining away mentioned above and serve as a definition of explaining away in the empirical research outlined in the present paper (see also Rehder & Waldmann, 2017; Rottman & Hastie, 2016).

1.2. Explaining away: Empirical account

Despite its ubiquity in human reasoning (Kelley, 1973; Pearl, 1988; Rottman & Hastie, 2016), empirical research on explaining away in the psychological sciences adopting the constrained definition outlined by the inequalities in (2) is somewhat limited and has insofar yielded mixed findings (for an overview see Rottman & Hastie, 2014).

Overall however, it appears that human explaining away inference, even in simple three-node common-effect causal structures (as

[1] Throughout the paper all random variables are binary: a random variable X (denoted by italicized letters) can take exactly one of the two values X or ¬X (denoted by non-italicized letters), where X indicates that X is present and ¬X indicates that X is absent.

in Figure 1), is fallible, thus emphasizing the significance of further investigating this evasive phenomenon.

Most of the studies exploring explaining away have reported that people explain away insufficiently or not at all (Davis & Rehder, 2017; Fernbach & Rehder, 2013; Liefgreen, Tesic & Lagnado, 2018; Morris & Larrick, 1995; Rehder & Waldmann, 2017; Rottman & Hastie, 2016; Sussman & Oppenheimer, 2011) or in some cases even display behaviour directly opposite to that of explaining away: $P(C_i|E, C_j) > P(C_i|E, \neg C_j)$ (Ferbach & Rehder, 2014; Rehder, 2014a) or $P(C_i|E, C_j) > P(C_i|E)$ (Rottman & Hastie, 2016, Experiment 1a).

1.3. Limitations of previous studies

Although the empirical studies on explaining away insofar speak to the robustness of people's deviation from the normative model, it is worth mentioning some limitations that are commonly found in these studies.

Firstly, the majority of studies neither conveyed to nor elicited from participants the prior probabilities of causes (see Rottman & Hastie, 2014), rendering any comparison to a normative model problematic. In most cases, priors indirectly dictate the amount of explaining away found in the normative model (see Morris & Larrick, 1995): lower priors imply a larger amount of explaining away than higher priors. As really high prior probabilities lead to minimal amounts of explaining away in the normative model, even if participants adopted the priors given to them and engaged in the correct pattern of inference, explaining away would most probably remain undetected. In the present work we : (i)provided participants with explicit priors and subsequently re-eliciting these to ensure they have been accepted, (ii) utilised low priors to maximise the amount of explaining away in the normative model and facilitate its detection and (iii) assigned different (low) priors to the two causes in the model to vary the amount of explaining away.

Secondly, the majority of studies exploring explaining away in common-effect structures report a violation of the Markov condition of independence, i.e. $P(C_i|C_j) \neq P(C_i|\neg C_j)$ (Rehder, 2014a, 2014b; Rehder & Burnett, 2005). In these cases, participants are misconstruing the two causes to be initially dependent, typically by assuming they are positively correlated. This is problematic since the higher the degree of positive correlation, the lower the normative amount of explaining away, with very high degrees of positive correlation potentially leading to a pattern opposite to explaining away (see Morris & Larrick, 1995). In order to guard ourselves against potential violations of the independence assumption, in the present work we: (i) explicitly stated that the two causes are independent, (ii) utilised cover stories that intuitively minimized participants' inclination to view the two causes as

unconditionally dependent, and (iii) utilised qualitative relational questions to investigate people's understanding of independence.

Finally, despite explaining away being a relational concept, the majority of empirical studies on explaining away elicit participants' belief estimates in isolation and do not investigate whether participants understand the relational nature of this pattern of reasoning. To rectify this issue, in the present work we complement quantitative questions asking for numerical probability estimates of, for example, $P(C_i|E, C_j)$, with qualitative relational questions asking them to consider whether $P(C_i|E, C_j)$ is less than, greater than, or equal to $P(C_i|E)$.

2. MOTIVATIONS FOR PRESENT WORK

In our previous study (Liefgreen et al., 2018), we tried to address some of the above-mentioned methodological issues often found in empirical studies on explaining away. Despite concluding that participants accepted priors of causes and did not violate the assumption of independence, we still observed insufficient explaining away. Moreover, a large cluster of participants did not update the probabilities of causes from their priors, given the presence of the effect or even given the presence of the effect and the other cause. This together with participants' explanations of the way they updated the probabilities led us to hypothesize that participants in this cluster may be interpreting probabilities differently, or more specifically, as propensities.

2.1. Probability interpretations

A large number of studies exploring human reasoning under uncertainty implicitly or explicitly assume the subjective probability interpretation where probabilities are identified as degrees of belief of a particular person about a certain event occurring. However, in philosophy of statistics one finds a whole spectrum of probability interpretations, one of which is the propensity interpretation (Popper, 1959; Giere, 1973). According to this interpretation probabilities are propensities (or tendencies and dispositions) of a particular physical system to produce an outcome (Hajek, 2012). For example, the statement that the probability of a coin to land Heads equals ½ is equivalent to the statement that there is a coin tossing set-up and that on a particular trial the strength of the propensity for this coin to land Heads is ½. This propensity is objective, it is part of the physical world, and it does not depend on our subjective beliefs about the coin landing Heads.

How does this relate to explaining away? Imagine a situation where there are two coins tossed at the same time, each with a coin bias

of $\frac{1}{5}$ for Heads. In this set-up there is also a light bulb that will turn on if at least one coin lands Heads. Here, it is perfectly natural to ask about the propensity for the light bulb to turn on if Coin 1 landed Heads, i.e. $P(E|C_1)$. However, the propensity of Coin 1 to have landed Heads given that the light bulb turned on is simply the original propensity for Coin 1 to land Heads: whether or not the light bulb turns on does not affect the propensity/the coin bias of Coin 1 to land heads, i.e. $P(C_1|E) = P(C_1) = \frac{1}{5}$.[2] In the same vein, according to the propensity interpretation, observing the effect (or another cause) would not change the propensity of the cause in question to happen. This implies that people who interpret probabilities as propensities in explaining away situations will violate the normative account by not updating their estimates given the presence of the effect, or the presence of the alternative cause (i.e. they would be repeating the priors).

This behaviour could therefore be partly driving the insufficiency observed in empirical studies of explaining away as repeating the priors would drive the average sample estimate away from the normative one. This seems increasingly plausible in light of the psychology literature suggesting that people may be able to distinguish between different variants of uncertainty, one of which is propensity (see Fox & Ülkümen, 2011; Kahneman & Tversky, 1982), and studies suggesting that people are sensitive to different probability interpretations (Ülkümen, Fox, & Malle, 2016) and may in fact be thinking of probabilities as propensities (Keren & Teigen, 2001).

3. EXPERIMENT OVERVIEW

The main aim of the present experiment was to empirically test whether propensity interpretations of probability partly drive the observed deviation of people's explaining away inferences from the normative ones. We adopted a novel experimental design that addressed the methodological confounds employed by previous studies and manipulated the properties of cover stories within which we embedded our CBN. In our experiment, all participants were required to reason with the same three-node common-effect structure depicted in Figure 1, parameterized such that causes had unequal low priors ($P(C_1) = 0.2$ and $P(C_2) = 0.1$) to increase the normative amount of explaining away in the model. Moreover, we utilised a deterministic setup wherein the

[2] This intuition has been (formally) outlined in Humphreys (1985), who employs it to argue that propensities are inconsistent with probabilities. This inconsistency is commonly known as 'Humphreys' paradox' in the literature.

presence of one cause entailed the presence of the effect ($P(E|C_1, C_2) = P(E|C_i, \neg C_j) = 1$, and the absence of both causes entailed the absence of the effect ($P(E|\neg C_1, \neg C_2) = 0$.

To test whether the propensity interpretation affected people's judgements on inferences relating to independence of causes, diagnostic reasoning, and explaining away, we manipulated properties of cover stories that, to a larger or a lesser extent, accentuated the propensity interpretation.

3.1. Manipulating cover stories

We embedded our common-effect structure within three different cover stories: one involving coin-tossing, one involving balls and containers, and one involving a dinner party.

In the coin-tossing cover story, the two causes (C_1 and C_2) were represented by two coins (binary variables; either Heads or Tails) tossed with the probability p_i for Heads by two coin-tossing mechanisms in separate rooms. If at least one coin landed Heads, a light bulb (common effect) stored in a different unit would switch on. From the propensity interpretation point of view, p_i is the propensity for a coin to land Heads given a coin-tossing set-up and that propensity does not change whether or not the light bulb (i.e. the effect) is on or off. As the questionnaire prompted participants to answer diagnostic reasoning and explaining away questions pertaining to the coins (see Section 4 below), we argue that the propensity interpretation would be strongly pronounced in this scenario.

In the balls and containers cover story the two causes were represented by two balls (binary variables; either copper or rubber) randomly selected from independent containers and placed on two gaps in an electric circuit. If at least one of the two balls was copper, a light bulb in the circuit (common effect) would turn on. Here, we follow Giere (1973) in arguing that the propensity is still present in this set-up, but it is at the level of a random sampling mechanism, not at the level of balls. As we prompted participants to questions pertaining to the balls and not to the random sampling mechanism, we argue that the propensity interpretation is less pronounced in this cover story compared to the coin-tossing one.

Finally, in the dinner party cover story the two causes were represented by two individuals, Michael and Tom, and the common effect was represented by a third individual, Helen, who would drink wine only if at least one of the two aforementioned people brought wine to a dinner party ('Helen' was a binary variable; either 'drinking wine' or 'not drinking wine'). In this set-up, the probability p_i of whether a person brings wine to the party was determined purely by host's subjective

estimates, implying that in this scenario the propensity interpretation is the least pronounced (if at all present).

Given the above rationale, we predicted that the proportion of participants whose reasoning aligns with the propensity interpretation, i.e. who would respond $P(C_i|E) = P(C_i|E, C_j)$ would be the highest when reasoning with the coin-tossing cover story, smallest when reasoning with the dinner party cover story, and fall in between these when reasoning with the ball containers cover story.

4. METHODS

4.1. Participants and design

A total of 271 participants (N_{MALE}= 111, 4 identified their gender as other; M_{AGE} = 32.2 years, SD = 10.2) were recruited from Prolific Academic (www.prolific.ac). All participants were native English speakers who gave informed consent and were paid £1 for partaking in the present study, which took on average 10.8 minutes (SD= 5.4) to complete. Eight participants were excluded as they did not pass the attention check, leaving a total of 263 participants in the analyses.

A between-subjects design was employed, and participants were randomly allocated to one of the three groups which differed in the cover story they were required to reason with. Group 1 (n=87) was presented with the coin-tossing cover story; Group 2 (n=87) with the ball containers cover story and Group 3 (n = 89) with the dinner party cover story.

4.2. Materials

Each of the three groups was asked to complete the same inference questionnaire ($N_{QUESTIONS}$ = 12) comprising of the questions outlined in Table 1. Although all participants completed this inference questionnaire, as mentioned in Section 4.1, participants in each group were required to reason with different cover stories, within which we embedded the common-effect structure. For cover story details see Section 3.1 and for full details on the inference questionnaire visit Open Science Framework, https://osf.io/zm6ec/.

Table 1: Inference types and questions found in questionnaire.

Question number	Inference Type	Key Inferences	Question Type
1	Priors	$P(C_1)$	Quantitative
2		$P(C_2)$	Quantitative
3	Independence	$P(C_2\|C_1)$	Qualitative

4		$P(C_1 \mid \neg C_2)$	Qualitative
5,6	Diagnostic Reasoning	$P(C_1 \mid E)$	Qual. + Quant.
7,8		$P(C_2 \mid E)$	Qual. + Quant.
9, 10	Explaining Away	$P(C_1 \mid E, C_2)$	Qual. + Quant.
11,12	Logic[3]	$P(C_1 \mid E, \neg C_2)$	Qual. + Quant.

4.3. Procedure

Participants in each of the three groups were initially presented with the pertinent cover story and were given explicit information on the common-effect model embedded within the cover story including the prior probability of each cause, and the causal relationships within the model. This was done in both textual form and through a graphical representation. Participants were provided with a textual account by which each cause could independently bring about the common effect. Subsequently, participants were presented with the inference questionnaire (for questions and associated inferences see Table 1).

Questions marked as quantitative in Table 1 required participants to provide numerical estimates on a slider (scale of 0-100 %). Questions marked as qualitative required participants to select one of three options: the probability increases, decreases, or stays the same when asked about e.g. $P(C_2 \mid C_1)$ given no knowledge about the state of E. To investigate participants' diagnostic and explaining away reasoning we employed both qualitative and quantitative question formats. This enabled us to capture the relational nature of explaining away, and retain focus on the direction and magnitude of change of beliefs given certain evidence. Additionally, in order to better understand participants' reasoning, some questions prompted participants to provide written explanations for their answers.

5. RESULTS

Participants' answers to all quantitative questions in the inference questionnaire are graphically represented in Figure 2. The results section will be sub-divided by analyses carried out for each inference type.

[3] We have labelled questions 11 and 12 as 'logic' questions, since our set-up is deterministic and learning that one cause did not happen, whilst knowing that the effect happened, entails (by logic) that the other cause must have happened, i.e. $P(C_1 \mid E, \neg C_2) = 1$.

5.1. Prior probabilities and independence of causes

Within each group we obtained the percentage of people who correctly answered[4] both questions on prior probabilities of causes (Q1 and Q2 in Table 1). Within Group 1 this was 88.5% of participants, within Group 2, 77% and within Group 3, 72%. Additionally, we obtained the percentage of people who correctly answered *both* questions regarding the independence of causes (Q3 and Q4 in Table 1). Within Group 1, this was 88.5%; within Group 2, 87.4%; and within Group 3, 91%. These high percentages illustrate that overall participants accepted the priors of causes that the experimenters explicitly stated and correctly regarded the causes as initially independent in all groups.

5.2. Logic

Independent analyses were conducted on qualitative and quantitative 'logic' questions (Q11 and Q12 in Table 1).

5.2.1. Qualitative

The percentage of participants who correctly answered the qualitative logic question was 62.1% in Group 1, 81.6% in Group 2 and 74.3% in Group 3. A Chi-Square test of independence illustrated these proportions significantly differed, $\chi^2\ (2) = 8.5, p = 0.001$. Bonferroni corrected post-hoc pairwise comparisons illustrated the only significant difference to be between the proportions of Group 2 and Group 3, $p = 0.004$.

5.2.2. Quantitative

The percentage of participants who correctly answered the quantitative logic question was 67.8% in Group 1, 81.6% in Group 2 and 70.8% in Group 3. A Chi-Square test of independence illustrated no significant difference in these proportions, $\chi^2\ (2) = 4.7, p = 0.09$. These results suggest that participants correctly understood the deterministic set-up of our experiment, in contrast to, for instance, Rottman and Hastie (2016).

[4] Answers to quantitative questions were codes as correct (1) if they were ±2% of normative answers, otherwise they were coded as incorrect (0).

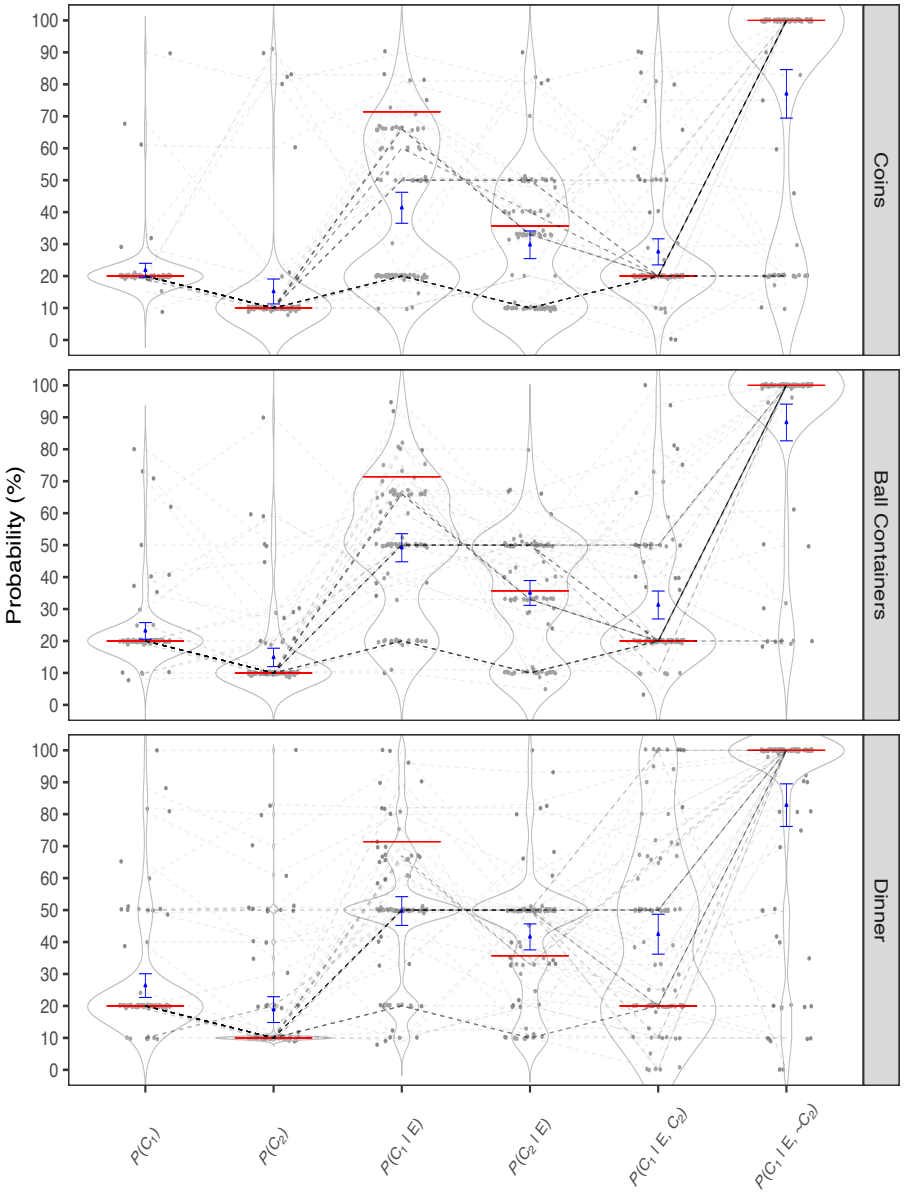

Figure 2: Participants' responses to quantitative questions. Red lines are normative answers. Blue dots are empirical averages with 95% confidence intervals as error bars. Dotted lines depict how participants changed their probability estimates from one questions to another, with darker lines indicating more participants changing the probabilities in the same way

5.3. Explaining away: Relational concept

Given the relational nature of explaining away, to better investigate participants' updating behaviour across this pattern of inference, we conducted aggregate analyses on questions pertaining to diagnostic reasoning, explaining away and logic (Q5-12 in Table 1). Independent analyses were conducted on qualitative and quantitative relational explaining away questions.

5.3.1. Qualitative

Given that we required participants to make two qualitative diagnostic reasoning inferences, i.e. $P(C_1|E)$ and $P(C_2|E)$, if a participant answered both questions regarding qualitative diagnostic reasoning correctly, we coded the response as 1; otherwise 0.

The proportion of participants who correctly answered all qualitative questions pertaining to relational explaining away was 28.7% in Group 1, 33.3% in Group 2 and 21.3% in Group 3. A Chi-Square test of independence illustrated no significant difference between these proportions,$\chi^2\ (2) = 3.2, p = 0.2$ suggesting that participants across all conditions performed relatively poorly on qualitative relational explaining away.

5.3.2. Quantitative

In regard to quantitative relational explaining away, we analysed questions relating to the updating o fC_1, namely $P(C_1|E)$, $P(C_1|E, C_2)$ and $P(C_1|E, \neg C_2)$.

A repeated-measures ANOVA with a Greenhouse Geisser correction was carried out on the average probability estimates on the relational explaining away questions, within each of the groups (see Figure 3). Results illustrated a significant difference between these estimates within Group 1; $F(1.59,122.6) = 95.6, p < 0.0001$, within Group 2; $F(1.8,140) = 167.5, p < 0.0001$ and within Group 3;$F(1.6,126) = 57, p < 0.0001$. Post-hoc paired t-tests allowed us to obtain the 95% confidence intervals (CI) of the difference in participants' average probability estimates between pairs of inferences of interest (see Table 2 below).

Since participants in all groups under-adjusted their belief estimates (i.e. the normative difference was not included in any of the 95% CI of the empirical differences) we were able to conclude that there was insufficient explaining away in all groups.

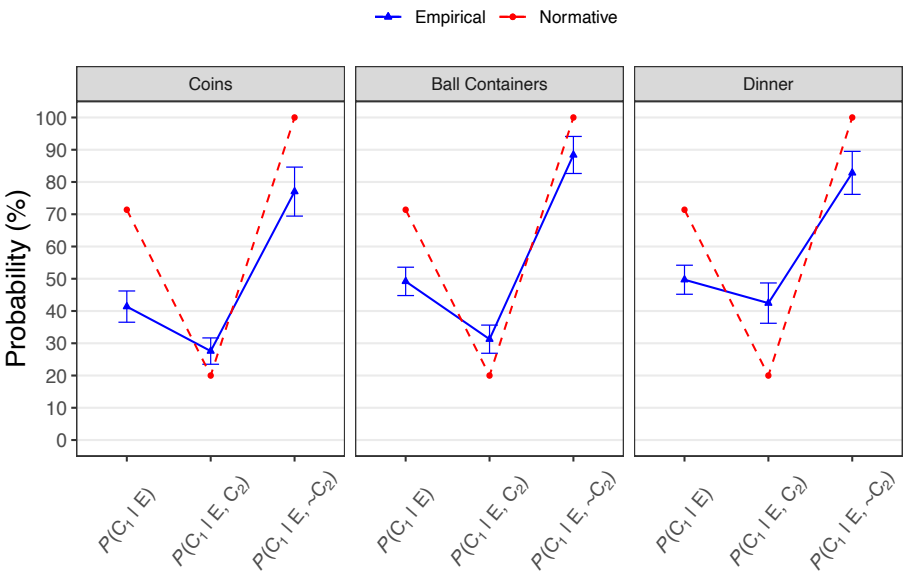

Figure 3: Participants' quantitative relational explaining away responses. Error bars are 95% confidence intervals.

Table 2: Within group explaining away.

Inferences	Normative difference	Empirical difference	95% CI of empirical difference
Group 1			
A-B	36	13.8	[8.5, 19]
C-B	80	49.4	[40.7, 58.2]
Group 2			
A-B	36	17.9	[12.6, 23.3]
C-B	80	57.1	[49.8, 64.4]
Group 3			
A-B	36	7.2	[1.5,13]
C-B	80	40.4	[31.9,48.9]

Note: A: = $P(C_1|E)$, $B{:=}\ P(C_1|E,C_2)$, C $:=\ P(C_1|E,\neg C_2)$.

5.4. 'Stay the same'

To test the hypothesis that participants in certain groups would be more prone to interpret probabilities as propensities, we obtained the proportions of participants in each group who did not update from their priors in diagnostic reasoning and explaining away questions (Q5-10 in Table 1). Independent analyses were conducted on qualitative and quantitative inferences in this section.

5.4.1. Qualitative

We computed the proportion of participants who selected the 'stay the same' option to both diagnostic reasoning and to explaining away qualitative questions, i.e. after being asked about $P(C_1|\,E)$, $P(C_2|\,E)$ and $\mathrm{P}(C_1|E, C_2)$ (see Figure 4). In Group 1, this was 44.8%, in Group 2, 33.3% and in Group 3, 23.6%. A Chi-Square test of independence illustrated that these proportions significantly differed, $\chi^2\,(2) = 8.87, p = 0.013$. Bonferroni corrected ($\alpha = 0.017$) post-hoc pairwise comparisons showed the only significant difference to be between the proportions of Group 1 and Group 3, $p = 0.005$.

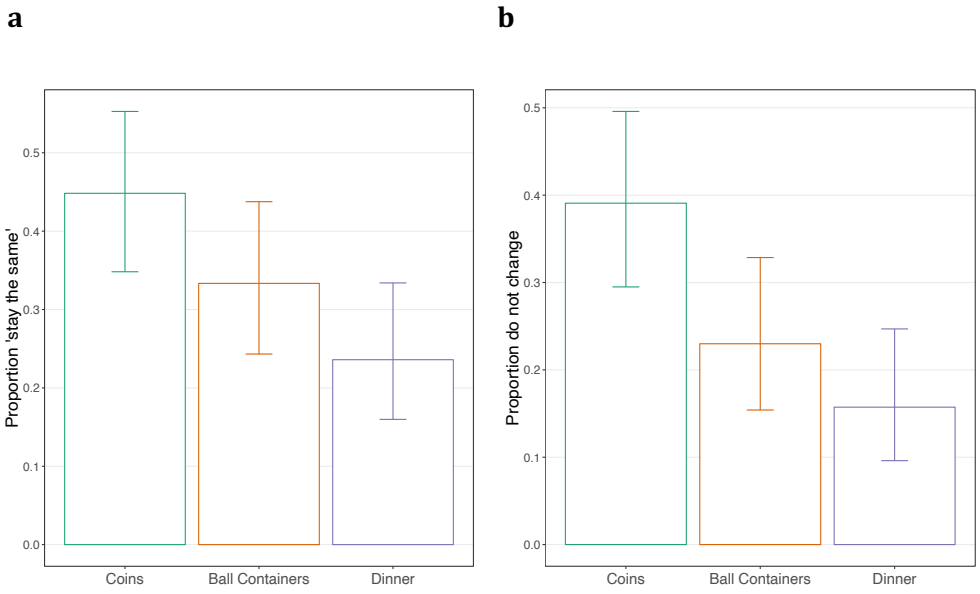

Figure 4. (a). The proportions of participants who chose 'stay the same' option on qualitative Q5, Q7, and Q9 and (b) who did not change their estimates in Q6, Q8, and Q10 compared to their own stated priors. Error bars are 95% confidence intervals.

5.4.2. Quantitative

We obtained the proportion of participants who did not update from their own stated priors on both diagnostic reasoning and explaining away quantitative questions (see Figure 5). In Group 1, this was 39.1% of participants, in Group 2, 23%, and in Group 3, 15.7%. A Chi-Square test of independence illustrated these proportions significantly differed, $\chi^2\ (2) = 13.07, p = 0.0014$ post-hoc pair-wise comparisons (α = 0.017) showed the only significant difference to be between the proportions of Group 1 and Group 3, p = 0.0009.

Overall, these results are in support of our hypothesis as Group 1 'stayed the same' significantly more, in both qualitative and quantitative inferences, compared to Group 3, with Group 2 falling in between.

5.5. Explaining away: excluding 'stay the same'

In this section we explored whether participants interpreting probabilities as propensities accounted for the observed insufficiency in explaining away. For qualitative inferences, we ran the same analysis on explaining away as in Section 5.4, but removing the set of participants who answered 'stay the same' to all qualitative questions relating to $P(C_1|\ E)$, $P(C_2|\ E)$, and $\mathrm{P}(C_1|E, C_2)$. For quantitative inferences we removed the set of participants who did not change their quantitative explaining away estimates regarding $P(C_1|\ E)$, $P(C_2|\ E)$ and $\mathrm{P}(C_1|E, C_2)$ compared to their stated priors.

5.5.1. Qualitative

Within the new subset, the proportion of participants who correctly answered all qualitative relational explaining away questions was 52.1% in Group 1, 50% in Group 2 and 27.9% in Group 3. A Chi-Square test of independence illustrated a significant difference between these proportions, $\chi^2\ (2) = 9, p = 0.01$. Bonferroni corrected (α = 0.017) post-hoc pairwise comparisons showed a significant difference between Group 1 and Group 3, p = 0.008 and between Group 2 and Group 3, p = 0.011. These percentages are notably higher than those reported in Section 5.4.1, suggesting that participants who interpreted probabilities as propensities were mostly driving the insufficiency in qualitative explaining away.

5.5.2. Quantitative

Similarly, within the new subset of data, a repeated-measures ANOVA with a Greenhouse Geisser correction was carried out on the average

probability estimates on the relational explaining away questions, within each of the groups. Results illustrated a significant difference between these estimates within Group 1; $F(1.58, 82.6) = 70.43, p < 0.0001$, within Group 2; $F(1.8,116.4) = 141.5, p < 0.0001$ and within Group 3; $F(1.6,119.4) = 63, p < 0.0001$. Post-hoc pairwise t-tests allowed us to obtain 95% confidence intervals (CI) of the difference in the average empirical probability estimates between pairs of inferences of interest (see Table 3 below). This analysis illustrated that there was no sufficient explaining away in any group, since the normative difference was not included in any of the 95% CI of the empirical differences.

Notably, however, compared to findings reported in Section 5.4.2, the insufficiency was less pronounced as the 95% CI of the empirical differences were now closer to the normative differences. This suggests that participants interpreting probabilities as propensities were significantly contributing to the observed insufficiency of quantitative relational explaining away.

Table 3: Within group explaining away excluding participants who did not change their quantitative estimates from their stated priors.

Inferences	Normative difference	Empirical difference	95% CI of empirical difference
Group 1			
A-B	36	22.7	[14.9, 30.4]
C-B	80	53.9	[42.7, 65.2]
Group 2			
A-B	36	23.2	[16.9, 29.6]
C-B	80	58.7	[50.47, 66.9]
Group 3			
A-B	36	8.6	[1.8, 15.4]
C-B	80	41.5	[31.9,50.9]

Note: A: = $P(C_1|E), B: = P(C_1|E, C_2)$, C ≔ $P(C_1|E, \neg C_2)$.

6. DISCUSSION

Over the past few decades, causal Bayesian networks have been successfully employed in many domains of human reasoning. Despite this, empirical work in the psychological literature has repeatedly illustrated that when engaging in explaining away, people violate the normative CBN model in numerous ways. One recurrently reported

violation in empirical studies of explaining away pertains to the violation of the Markov assumption of independence (Mayrhofer & Waldmann, 2015; Rehder & Waldmann, 2017; Rottman & Hastie, 2016). Another pertains to people's under-adjustment of probabilities and insufficient explaining away (Davis & Rehder, 2017; Fernbach & Rehder, 2013; Liefgreen et al., 2018; Morris & Larrick, 1995; Rehder & Waldmann, 2017; Rottman & Hastie, 2016; Sussman & Oppenheimer, 2011). Although this insufficiency has partly been attributed to structural violations of the normative model such as the assumption of independence, we argue it instead to be the product of methodological confounds of previous studies and participants interpreting probabilities as propensities.

In the present study, we utilised a novel methodology to address the issues often found in empirical studies of explaining away. For example, we explicitly stated priors and re-elicited these from participants and we utilised relational qualitative and quantitative questioning. This approach was seemingly successful in making participants understand the parameters and relational properties found within the common-effect structure they were required to reason with. As such, in all three conditions a high proportion of participants correctly answered questions regarding prior probabilities of causes, independence of causes, and the final logic question. This allowed us to conclude that the assumption of independence remained intact in all conditions, participants had accepted the priors given to them, and they understood what circumstances were necessary to bring about the common effect. This is in contrast to the vast majority of findings reported by studies in the extant literature (e.g. Rottman & Hastie, 2016) and allowed us to make meaningful comparisons between people's inferences and those dictated by the normative model.

Despite these encouraging improvements, our main findings echo those of the extant literature, including our previous study (Liefgreen et al., 2018), as participants systematically violated the normative account of explaining away by under-adjusting their belief estimates in all three conditions. Pitfalls in relational explaining away were visible at the level of both diagnostic reasoning and explaining away. As such, participants in all groups performed extremely poorly in qualitative and quantitative diagnostic reasoning questions and slightly better, but still sub-optimally, in questions relating to explaining away. In our study, deviations from the normative model could not be attributed to violations of the independence assumtion, but instead seem to arise, at least in part, from some participants interpreting probabilities as propensities. As predicted by our propensity interpretation hypothesis, a larger number of participants engaged in this 'stay the same' behaviour on both qualitative and quantitative relational explaining away inferences in Group 1 reasoning with the coin-tossing cover story than in Group 3,

reasoning with the social dinner party cover story, with Group 2 reasoning with balls and containers cover story falling in between. Further analyses on people's qualitative and quantitative relational explaining away without the cohort of participants who seemed to have interpreted probabilities as propensities showed that participants' insufficiency was now noticeably less pronounced, suggesting that the cluster of participants who interpreted probabilities as propensities was, at least partly, driving the extreme insufficiency observed in the overall sample.

Another large cluster of participants' responses in our data that seemed to have been responsible for the overall insufficient explaining were those that updated their probabilities in diagnostic reasoning using an erroneous strategy whereby $P(C_1|E) + P(C_2|E) = 1$, with $P(C_1|E) = P(C_2|E) = 0.5$ (assigning equal probability to each cause), or $\mathrm{P}(C_1|E) = 0.67$ and $\mathrm{P}(C_2|E) = 0.33$ (assigning a probability to each cause that reflects the 2:1 priors ratio) (see Figure 2). Further experimental work is needed to test for the pervasiveness of this erroneous strategy in diagnostic reasoning within explaining away.

Overall, our findings advocate for future work that not only investigates whether people's explaining away reasoning differs from normative predictions, but, given the proven robust nature of the insufficiency of explaining away, also explores when and why these deviations occur. Moreover, in order to investigate people's differential interpretations of probability more directly, we suggest future research could manipulate the phrasing of questions (see Ülkümen et al., 2016) so that participants are being asked either about propensities or subjective probabilities when making inferences on independence, diagnostic reasoning and explaining away.

REFERENCES

Davis, Z., & Rehder, B. (2017). The causal sampler: A sampling approach to causal representation, reasoning, and learning. In *Proceedings of the cognitive science society.*

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*, 21–35. Oslo, Norway: Universitetsforlaget.

Giere, R. N. (1973). Objective single-case probabilities and the foundations of statistics. In *Studies in logic and the foundations of mathematics, 74*, 467–483

Griffiths, T. (2001). Explaining away and the discounting principle: Generalising a normative theory of attribution. *Unpublished manuscript.*

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.

Keren, G., & Teigen, K. H. (2001). The probability-outcome correspondence principle:A dispositional view of the interpretation of probability statements. *Memory & cognition, 29*(7), 1010–1021.

Liefgreen, A., Tešić, M., & Lagnado, D. (2018). Explaining away: significance of priors, diagnostic reasoning, and structural complexity. In *Proceedings of the 40th annual conference of the cognitive science society.* Austin: Texas.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science, 39*, 65–95.

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.

Neapolitan, R. E. (2003). *Learning Bayesian networks.* Pearson Prentice Hall UpperSaddle River, NJ.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kauffman.

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

Rehder, B. (2014b). The role of functional form in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 670–692.

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.

Sussman, A. B., & Oppenheimer, D. M. (2011). A causal model theory of judgment. In *Proceedings of the 33rd annual conference of the cognitive science society.* Austin: Texas.

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.