The propensity interpretation of probability and diagnostic split in explaining away[☆]

Marko Tešić[a,1,*], Alice Liefgreen[b,1], David Lagnado[b]

[a] *Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK*

[b] *Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK*

[☆] This research has been partly supported by the Leverhulme Trust under Grant RPG-2016-118 CAUSAL-DYNAMICS.

[*] Corresponding author

Email addresses: mtesic02@mail.bbk.ac.uk (Marko Tešić), alice.liefgreen.15@ucl.ac.uk (Alice Liefgreen), d.lagnado@ucl.ac.uk (David Lagnado)

[1] These two authors have equally contributed to the present research.

Abstract

Causal judgements in explaining-away situations, where multiple independent causes compete to account for a common effect, are ubiquitous in both everyday and specialised contexts. Despite their ubiquity, cognitive psychologists still struggle to understand how people reason in these contexts. Empirical studies have repeatedly found that people tend to 'insufficiently' explain away: that is, when one cause explains the presence of an effect, people do not sufficiently reduce the probability of other competing causes. However, the diverse accounts that researchers have proposed to explain this insufficiency suggest we are yet to find a compelling account of these results. In the current research we explored the novel possibility that insufficiency in explaining away is driven by: (i) some people interpreting probabilities as propensities, i.e. as tendencies of a physical system to produce an outcome and (ii) some people splitting the probability space among the causes in diagnostic reasoning, i.e. by following a strategy we call 'the diagnostic split'. We tested these two hypotheses by manipulating (a) the characteristics of cover stories to reflect different degrees to which the propensity interpretation of probability was pronounced, and (b) the prior probabilities of the causes which entailed different normative amounts of explaining away. Our results were in line with the extant literature as we found insufficient explaining away. However, we also found empirical support for our two hypotheses, suggesting that they are a driving force behind the reported insufficiency.

Keywords: Explaining away, Probability interpretation, Propensity, Causal Bayesian networks, Causal inference

1. Introduction

Every day we make numerous judgements and inferences that rely on our beliefs about how events or items of information are causally related to each other. For example, on the way to work people may think of possible causes that could lead them to be late to an important meeting such as heavy traffic, a broken elevator, or adverse weather conditions. The vast majority of these causal judgments occur *under uncertainty*.

Since erroneous causal probabilistic inferences, particularly in specialized contexts, can lead to deleterious consequences, understanding how these inferences are made is critical. Consider for instance a real-world 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. From her experience the social worker knows that bruises could also be the product of alternative independent causes, one of which is a rare blood disorder 'haemophilia'. Since she does not know for certain whether the child was physically abused and/or whether he suffers from haemophilia, but she knows of the presence of bruises, she increases the probability of each potential cause. If after a medical examination the social worker found out that the child definitely suffers from haemophilia, then the probability of the child being physically abused would decrease, since haemophilia is sufficient to explain the bruises. If on the other hand the medical examination revealed that the child definitely *does not* suffer from haemophilia, then the probability of the child being abused would further increase as a result.[2] This scenario illustrates a pattern of reasoning known as 'explaining

[2] The importance of understanding explaining away relationships in these contexts is clearly reflected in the American Academy of Pediatrics' (AAP) clinical report where conducting laboratory evaluations with the understanding that presence of a bleeding disorder does not rule out physical abuse is highly emphasized (Anderst, Carpenter, & Abshire, 2013). Furthermore, the AAP also warns physicians that inappropriate

away'.[3] In more general terms, explaining away describes a situation 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. Subsequently, after learning of the occurrence of one cause (the child suffers from haemophilia) the probability of the alternative cause(s) decreases (physical abuse). If, conversely, we learned that a cause did not happen (the child does not suffer from haemophilia), the probability of the other cause(s) further increases (physical abuse).

An increasingly popular approach in the cognitive science to modeling causal reasoning in general and explaining away in particular employs Causal Bayesian Networks (CBNs). We will first describe CBNs and a CBN model for explaining away and subsequently outline previous empirical work on explaining away in the psychological literature as well as the potential shortcomings of this work. Finally, we will discuss motivations and details of the experimental work presented in this paper.

1.1. Explaining away: normative account

Causal Bayesian networks (Pearl, 2009; Neapolitan, 2003) can be used to represent probabilistic knowledge in a graphical manner. They are directed acyclic graphs (DAGs) with nodes representing random variables[4] and arrows representing the causal relationships between these variables. Arrows in CBNs point only in one direction (directional graph) and following the arrows there is no path that starts and finishes at the same node (acyclic graph).

diagnostics of child abuse can lead to the potential prosecution of an innocent person.

[3] A related concept to explaining away is discounting. For the distinction between the two concepts see Khemlani and Oppenheimer (2011), Rehder and Waldmann (2017), Rottman and Hastie (2014).

[4] In this paper, all random variables in CBNs are binary: a random variable X (denoted by italicized letters) can take exactly two values X or ~X (denoted by non-italicized letters), where X indicates that X is present and ~X indicates that X is absent.

The computational machinery of CBNs grounded in the 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. However, in order to perform these calculations one needs to fully parameterize the CBNs by specifying (i) the prior probabilities (or priors) of all root nodes (i.e. nodes that do not have incoming arrows) and (ii) the conditional probabilities of each remaining node given all the values of their direct causes (i.e. nodes they are directly linked to).

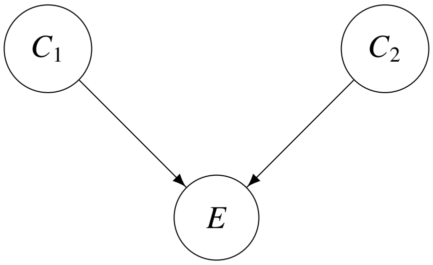

Figure 1: A CBN model of explaining away

Consider the graph in Figure 1, typically referred to as a common-effect CBN. It consists of three nodes representing three random variables: two causes, C_1 and C_2, and one common effect, E. The graph is directed and acyclic and since C_1 and C_2 are interpreted as causes and E as an effect, the arrows have a causal interpretation making the DAG a CBN. To fully parametrize this CBN, one needs to specify the prior probabilities of the two causes, i.e. $P(\mathrm{C}_1)$ and $P(\mathrm{C}_2)$, as well as the conditional probabilities of the effect E given the presence and/or absence of each cause, i.e. $P(\mathrm{E} \mid \mathrm{C}_1, \mathrm{C}_2)$, $P(\mathrm{E} \mid \mathrm{C}_1, {\sim}\mathrm{C}_2)$, $P(\mathrm{E} \mid {\sim}\mathrm{C}_1, \mathrm{C}_2)$, and $P(\mathrm{E} \mid {\sim}\mathrm{C}_1, {\sim}\mathrm{C}_2)$.[5] Once one specifies these parameters, one can compute for instance $P(\mathrm{C}_1 \mid \mathrm{E})$, or $P(\mathrm{E} \mid \mathrm{C}_2)$, or $P(\mathrm{C}_1, \mathrm{E} \mid {\sim}\mathrm{C}_2)$, etc.

[5] Since all variables are binary, one has it that $P({\sim}\mathrm{C}_1) = 1 - P(\mathrm{C}_1)$, $P({\sim}\mathrm{C}_2) = 1 - P(\mathrm{C}_2)$, $P({\sim}\mathrm{E} \mid \mathrm{C}_1, \mathrm{C}_2) = 1 - P(\mathrm{E} \mid \mathrm{C}_1, \mathrm{C}_2)$, etc.

In addition to quantitative computations, CBNs allow one to make qualitative inferences on the (un)conditional independencies and dependencies between variables in the network. For instance, the structure of the network in Figure 1 dictates that regardless of the network parametrization, C_1 and C_2 are unconditionally independent. This means that not knowing the state of the common effect variable E, learning that C_1 is present or absent does not affect the probability of C_2 being present or absent and vice versa (or $P(C_i \mid C_j) = P(C_i)$ where $i \in \{1, 2\}$ for any value of C_i and C_j). However, depending on the network parameterization, the two causes may become conditionally dependent on E, i.e. upon learning that E is either present or absent, the presence or absence of C_1 may affect the probability of C_2 being present or absent and vice versa.

Situations involving explaining away can be modeled utilising common-effect CBNs such as the one in Figure 1 (see Pearl, 1988, 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. Once we learn that the child has bruises on his body, we update the probabilities of the two causes via diagnostic reasoning (i.e. reasoning from effects to causes, see Meder & Mayrhofer, 2017). The fact that the child has bruises on his body, now renders the two causes conditionally dependent, since, as per explaining away, additionally learning that the child is suffering from haemophilia would change (decrease) the probability that the child has been physically abused. Common-effect CBNs, however, do not always lead to the pattern of explaining away where after observing the effect, additionally learning one cause decreases the probability of the other. This is only the case when CBNs are parameterized such that the following inequality holds (see Wellman & Henrion, 1993):

$$P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j})\, P(\mathrm{E} \mid {\sim}\mathrm{C_i}, {\sim}\mathrm{C_j}) < P(\mathrm{E} \mid \mathrm{C_i}, {\sim}\mathrm{C_j})\, P(\mathrm{E} \mid {\sim}\mathrm{C_i}, \mathrm{C_j}) \tag{1}$$

for $i, j \in \{1, 2\}$; or in words, the product of the probability of evidence knowing both causes are true and the probability of evidence knowing neither cause is true is strictly less than the product of evidence knowing only one cause is true and the other false and the probability of evidence knowing the other cause is true and the first one is false. From Inequality (1) it follows (see Morris & Larrick, 1995; Griffiths, 2001):

$$P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) < P(\mathrm{C_i} \mid \mathrm{E}) < P(\mathrm{C_i} \mid \mathrm{E}, {\sim}\mathrm{C_j}) \tag{2}$$

The inequalities in (2) accord 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).

It is often assumed (and empirical studies have been conducted with this assumption in mind) that explaining away situations hold when both causes are generative: the probability of evidence given a cause is greater than the prior probability of evidence (i.e. $P(\mathrm{E} \mid \mathrm{C_i}) > P(\mathrm{E})$) (Cheng, 1997). This is true, meaning that Inequality (1) (and hence the inequalities in (2)) holds if the causes are generative. However, it is also the case that Inequality (1) holds if both or one of the causes is inhibitory, i.e. when the probability of evidence given that cause is less than the prior probability of evidence or $P(\mathrm{E} \mid \mathrm{C_i}) < P(\mathrm{E})$.[6] For example, sneezing can be prevented by taking antihistamine drugs and/or by turning on an air filtration system. Learning that a person is sneezing will decrease the probability

[6] Here we are not claiming that if $P(\mathrm{E} \mid \mathrm{C_i}) > P(\mathrm{E})$ then the cause is generative and if $P(\mathrm{E} \mid \mathrm{C_i}) < P(\mathrm{E})$ then the cause is inhibitory, as the two events can be positively or negatively correlated without them being causally related. Rather, we are taking that if a cause is generative, then $P(\mathrm{E} \mid \mathrm{C_i}) > P(\mathrm{E})$ and if a cause is inhibitory then $P(\mathrm{E} \mid \mathrm{C_i}) < P(\mathrm{E})$.

of them taking antihistamine drugs and will decrease the probability that the air filtration system is on in the space they occupy (i.e. $P(\mathrm{E} \mid \mathrm{C_i}) < P(\mathrm{E})$ for both causes). However, additionally learning that a person is taking antihistamine drugs will further reduce the probability of the air filtration system being on, i.e. $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) < P(\mathrm{C_i} \mid \mathrm{E})$, since the probability of sneezing is lower when both the person is taking the antihistamine drugs and the air filtration system is on than when the person is just taking the antihistamine drugs but the air flirtation system is off. Conversely, if we instead learnt that the person is not taking the antihistamine drugs then probability of the air flirtation system being on will go back closer to its prior. In this case, $P(\mathrm{C_i} \mid \mathrm{E}) < P(\mathrm{C_i} \mid \mathrm{E}, \sim\mathrm{C_j})$ since the probability of sneezing is higher if the person is not taking the antihistamine drugs and the filtration system is off than if they are not taking the antihistamine drugs but the filtration system is on. More technical details on when Inequality (1) holds with regards to the generative/inhibitory nature of causes are presented in Appendix A.

Although the above are interesting considerations, in this paper we exclusively refer to, and focus on, generative causes.

1.2. Explaining away: empirical account

Explaining away is an ubiquitous pattern of inference, found in a wide range of contexts including social attribution, medical diagnosis and legal domains (Kelley, 1973; Pearl, 1988; Rottman & Hastie, 2016). In specialised contexts, as demonstrated by the aforementioned legal scenario, erroneous intercausal reasoning inferences may have detrimental consequences. Despite its ubiquity and importance in human reasoning, 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 (see 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; Morris & Larrick, 1995; Rehder & Waldmann, 2017; Rottman & Hastie, 2016; Sussman & Oppenheimer, 2011) and in some cases even display behaviour directly opposite to that of explaining away: $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) > P(\mathrm{C_i} \mid \mathrm{E}, {\sim}\mathrm{C_j})$ (Fernbach & Rehder, 2013; Rehder, 2014a) or $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) > P(\mathrm{C_i} \mid \mathrm{E})$ (Rottman & Hastie, 2016, Experiment 1a). Importantly, the insufficiency of explaining away remains robust across the different methodologies utilised by researchers. For example, Rottman and Hastie (2016) taught participants the statistical parameters of the variables in the common-effect structure through experience-based trials, complemented by written and graphical information. Disparately, Fernbach and Rehder (2013, Experiment 3) provided participants with explicit information on the structure in textual and graphical formats only. Finally, Rehder and Waldmann (2017) compared three different formatting methods to convey information to the participants: description-only (written description of the causal model, without communicating parameters), experience-only (data regarding the parameters presented in a tabular format without the causal structure), and description-experience (combination of the former two formats). Similarly, people's error-prone explaining away behaviour is seemingly persistent over different probability elicitation methods. Typically, studies have elicited probabilities from participants in the form of numerical estimates (Rottman & Hastie, 2016). Other methods that have been used include a verbal point scale or inference ratings (Fernbach & Rehder, 2013; Sussman & Oppenheimer, 2011) and qualitative forced choice responses in which participants are required to select which one of two situations is more likely to have a certain variable present, on the basis of the states of the other variables (Rehder, 2014a). Despite the use of different information presentation formats and belief elicitation methods, all of the above-mentioned studies reported insufficient explaining away.

1.3. *Limitations of previous studies*

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

1.3.1. *Prior probabilities of causes*

The majority of the studies neither convey nor elicit prior probabilities to participants (see Rottman & Hastie, 2014), making it difficult to compare participants' inferences to the normative model since it is unclear what prior probabilities participants assumed. In some cases, authors expected their participants to infer information on the priors of causes, but never elicited their estimates, therefore leaving unclear whether participants had accepted them (e.g. Rehder & Waldmann, 2017). Exceptions to this trend are the few studies that explicitly stated and subsequently elicited priors from participants (Liefgreen, Tešić, & Lagnado, 2018), or utilised participants' own prior probability estimates to calculate the normative benchmark probabilities pertaining to explaining away (Morris & Larrick, 1995).

The importance of adopting transparency when dealing with priors in empirical studies of explaining away also lies in the fact that priors in most cases directly dictate the amount of explaining away found in the normative model (see Morris & Larrick, 1995). Typically, lower priors imply a larger amount of explaining away than higher priors, since Δ_1 and Δ_2 are usually larger when the priors are lower than when they are higher, where $\Delta_1 = P(\mathrm{C_i} \mid \mathrm{E}) - P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ and $\Delta_2 = P(\mathrm{C_i} \mid \mathrm{E}, {\sim}\mathrm{C_j}) - P(\mathrm{C_i} \mid \mathrm{E})$. 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. This suggests that for the normative amount of explaining away in the model to be accurately computed (and thus for the comparisons to the normative model to be informative), it is crucial to know what priors are being utilised in experiments, both by participants and by experimenters. Although most studies have not taken these points into consideration, there are a few exceptions, which should encourage researchers to use similar approaches. For example, some authors manipulated the prior probabilities of causes to reflect different amounts of normative explaining away (e.g. Rottman & Hastie, 2016) and others purposefully utilised low priors in order to increase the amount of explaining away in their normative model (e.g. Rehder & Waldmann, 2017).

In the present work we address these issues by (i) providing participants with explicit priors and subsequently re-eliciting these to ensure they have been accepted and (ii) assigning different priors ranging from low to high to the causes in the model to vary the normative amount of explaining away.

1.3.2. Independence of causes

A second matter that could be contributing to the pervasive insufficiency of explaining away pertains to the reported systematic violation of the condition of independence in studies exploring explaining away in common-effect structures, i.e. $P(\mathrm{C_i} \mid \mathrm{C_j}) \neq P(\mathrm{C_i} \mid {\sim}\mathrm{C_j})$ (Rehder, 2011, 2014a, 2014b; Rehder & Burnett, 2005; Rehder & Waldmann, 2017, Description-only condition; Rottman & Hastie, 2016, Experiment 1b). In these cases, participants seem to be regarding the two causes to be initially dependent, typically reporting a positive correlation between them. Now, a positive correlation between the causes would significantly lower the amount of explaining away in the normative model. Generally, 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). This then suggests that an insufficiency in explaining away could be explained by participants understating causes to be positively correlated in studies where positive correlation between the causes is found. What is more, in instances in which the causes are positively correlated, it may even seem intuitive to not reduce or minimally reduce the probability of one causes given the other, after observing the effect (see Morris & Larrick, 1995). To slightly modify our example, haemophilia and internal bleeding can both be causes of bruises on a body, but haemophilia and internal bleeding are also positively correlated: a person suffering from haemophilia is more likely to have internal bleeding even before knowing anything about bruises. So, when a doctor learns that a patient has bruises, additionally learning that the patient has internal bleeding would incur minimal to no reduction in the likelihood that the patient is suffering from haemophilia. This notion is empirically supported by a study of Morris and Larrick (1995), in which participants explained away significantly less in the condition in which they were communicated that the causes were positively correlated than in conditions in which the causes were said to be independent or negatively correlated.

Empirically detecting explaining away is, then, potentially particularly difficult in studies where participants report positive correlations between the causes. For instance, in Rottman and Hastie (2016) Experiment 1b, participants' average estimates relating to independence of the causes were $P(\mathrm{C_i} \mid \mathrm{C_j}) = .45$ and $P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}) = .35$ (see Table 5 in Rottman & Hastie, 2016), suggesting a posiive correlation between the causes and a violation of the independence assumption. If one, however, includes these participants' average estimates as parameters in the normative model instead of those stated in the study (i.e. $P(\mathrm{C_i} \mid \mathrm{C_j}) = .25$ and $P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}) = .25$), one gets that $P(\mathrm{C_i} \mid \mathrm{E}) = .54$ and $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) = .55$ (see Appendix B for more details). So, given the participants' reported positive correlation between the causes, the difference between $P(\mathrm{C_i} \mid \mathrm{E})$ and $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ is now negligible and slightly goes in the opposite direction to explaining away. Furthermore, these new normative probability values for $P(\mathrm{C_i} \mid \mathrm{E})$ and $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ closely approximate average participants' estimates: $P(\mathrm{C_i} \mid \mathrm{E}) = .58$ and $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) = .56$ (see Table 7 in Rottman & Hastie, 2016). This is in line with the study by Morris and Larrick (1995) and highlights the importance of ensuring that participants understand the independence relations between the causes in order to increase chances of detecting explaining away and make more direct comparisons to the normative model which is assumed by the experimenters and communicated to the participants. In our studies we seek to guard from potential violations of independence by (i) explicitly emphasizing, in both textual and graphical formats, that the two causes are independent, (ii) employing cover stories that intuitively would minimize participants' inclination to view the two causes as unconditionally dependent, and (iii) asking participants qualitative relational questions (see below) prompting them to compare the probability of C_i given the presence/absence of C_j (when the state of the effect E is unknown) to the prior probability of C_i.

1.3.3. Probability elicitation methods

A third factor that may be contributing to the reported insufficiency of explaining away in the psychological literature pertains to how belief updates are elicited from participants. Foremost, explaining away is a *relational* concept. In our previous example scenario, a social worker reduces the probability that the child has been physically abused upon learning that he is suffering from haemophilia *relative to* the probability that the child has been physically abused when it was unknown whether the child is suffering from haemophilia. Similarly, the social worker increases the probability that the child has been physically abused upon learning that he is *not* suffering from haemophilia *relative to* the probability that the child has been physically abused when it was unknown whether the child is suffering from haemophilia. This relational property of explaining away is more formally expressed in the inequalities in (2). It is then important to empirically explore whether people understand this relational nature of explaining away.

Most studies on explaining away elicit participants' belief estimates in isolation without asking participants to compare their estimates or rates to their other estimates or rates. For instance, participants are often required to provide an estimate of the probability of a cause given the presence of both the effect and another cause, i.e. $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$, but they are seldom asked also to consider the relation and direction of change of this probability compared to the probability of the cause given just the effect, i.e. $P(\mathrm{C_i} \mid \mathrm{E})$.

Despite the intuitive importance of asking qualitative relational questions when testing for explaining away, to the best of our knowledge only two studies have employed such or similar methods: Rehder (2014a) and Liefgreen et al. (2018). Our current research builds on these studies and we complement quantitative questions asking for numerical probability estimates of, for example, $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$, with qualitative relational questions asking them to consider whether $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ is less than, greater than, or equal to $P(\mathrm{C_i} \mid \mathrm{E})$. Further, we distinguish between *direct* explaining away which corresponds to what is usually referred to as an explaining away question, namely a question about $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$, of course in relation to $P(\mathrm{C_i} \mid \mathrm{E})$ (see for example Morris & Larrick, 1995) and explaining away as a *relational* concept captured by inequalities in (2) which includes the question about $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$, but also about $P(\mathrm{C_i} \mid \mathrm{E})$ and $P(\mathrm{C_i} \mid \mathrm{E}, \sim\mathrm{C_j})$ (see for example Rehder & Waldmann, 2017). This will allow us to present a more comprehensive view regarding explaining away.

2. Motivations for present work

Due to the potential methodological confounds mentioned above and the mixed findings of the extant empirical work on explaining away, we conducted an initial study to evaluate people's explaining away inferences (see Liefgreen et al., 2018) utilising a novel design. Despite concluding that participants accepted priors of causes and did not violate the assumption of independence, Liefgreen et al. (2018) still observed insufficient explaining away. A closer inspection of the data strongly suggested that participants' behaviour could be categorised into two clusters: (1) those who, in answering diagnostic reasoning questions (i.e. $P(\mathrm{C_i} \mid \mathrm{E})$), split the probability space between the two causes and answered such that $P(\mathrm{C_1} \mid \mathrm{E}) + P(\mathrm{C_2} \mid \mathrm{E}) = 1$ and (2) those who 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: $P(\mathrm{C_i}) = P(\mathrm{C_i} \mid \mathrm{E}) = P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$. The explanations participants in cluster (2) provided led us to hypothesize that these participants may be interpreting probabilities in a certain way, which has been referred to in the philosophical literature as 'propensities'.

The two conjectures regarding the two clusters prompted us to design the current study in which we not only aimed to address the limitations of previous studies by employing a novel experimental design (see Methods section), but we also attempted to test (i) whether people employ a strategy that we call 'the diagnostic split' in tackling diagnostic reasoning questions and (ii) whether a specific interpretation of probability partly drives the observed deviation of people's explaining away inferences from the normative ones. We will now describe the two hypotheses in more detail and outline how we will empirically address them.

2.1. Diagnostic split strategy

Experimental data from our previous study (Liefgreen et al., 2018) indicated that a significant number of participants provided answers to the diagnostic reasoning questions such that $P(\mathrm{C_1} \mid \mathrm{E})$ and $P(\mathrm{C_2} \mid \mathrm{E})$ added up to 1. This was particularly striking in the condition in which the stated prior probabilities were low, $P(\mathrm{C_1}) = .2$ and $P(\mathrm{C_2}) = .1$. In this condition, a number of participants either said $P(\mathrm{C_1} \mid \mathrm{E}) = P(\mathrm{C_2} \mid \mathrm{E}) = .5$ or provided a more sophisticated answer to reflect the $2 : 1$ ratio of the priors, i.e. $P(\mathrm{C_1} \mid \mathrm{E}) = .67$ and $P(\mathrm{C_2} \mid \mathrm{E}) = .33$ (the normative answers were $P(\mathrm{C_1} \mid \mathrm{E}) = .71$ and $P(\mathrm{C_2} \mid \mathrm{E}) = .36$). Participants' verbal reasoning explanations regarding $P(\mathrm{C_i} \mid \mathrm{E})$ questions suggested that they correctly believed that since the effect was observed one of the causes must have occurred, but incorrectly believed that as there are two causes, there is a .5 probability that either cause happened.[7] Other explanations suggested participants reasoned in the following way: Cause 1 is 20% likely to be happen, while Cause 2 is only 10% likely to happen, and as we know one of them happened, it is twice as likely to be Cause 1, so the probability that the Cause 1 happened is .67, while this is .33 for Cause 2. This led us to hypothesize that when engaging in diagnostic reasoning in cases where the two (or more) independent causes become exhaustive upon learning evidence, i.e. $P(\mathrm{C_1} \vee \mathrm{C_2} \vee \ldots \vee \mathrm{C_n} \mid \mathrm{E}) = 1$ since $P(\mathrm{E} \mid {\sim}\mathrm{C_1}, {\sim}\mathrm{C_2}, \ldots, {\sim}\mathrm{C_n}) = 0$, but crucially they *do not* become mutually exclusive, i.e. $P(\mathrm{C_1}, \mathrm{C_2}, \ldots, \mathrm{C_n} \mid \mathrm{E}) \neq 0$ since $P(\mathrm{E} \mid \mathrm{C_1}, \mathrm{C_2}, \ldots, \mathrm{C_n}) > 0$, some people simply split the probability space between the two causes and assign each cause a .5 probability *when the causes had equal priors*. We dubbed this strategy 'the diagnostic split'.

It is worth noting the relationship between the diagnostic split strategy and the normative reasoning. Namely, as the priors of causes converge to 0, the normative diagnostic inferences approach to the the diagnostic split strategy.[8] Moreover, when the priors of the two causes follow a particular ratio, $a : b$, then, given priors are very close to 0, it normatively follows that $P(\mathrm{C_1} \mid \mathrm{E}) + P(\mathrm{C_2} \mid \mathrm{E}) \approx 1$ and $P(\mathrm{C_1} \mid \mathrm{E}) \approx \dfrac{a}{a+b}$ and $P(\mathrm{C_2} \mid \mathrm{E}) \approx \dfrac{b}{a+b}$ which follows the diagnostic split predictions (see Figure 2). As such, the diagnostic split hypothesis has its normative underpinnings and could be understood as an extreme approximation of the normative diagnostic reasoning.

Other empirical studies seems also to have found trends corresponding to the diagnos-

[7] The experimental design from our 2018 study was, like the experimental designs from Experiment 1 and 2 below, fully deterministic, i.e. $P(\mathrm{E} \mid \mathrm{C_1}, \mathrm{C_2}) = P(\mathrm{E} \mid \mathrm{C_i}, {\sim}\mathrm{C_j}) = 1$, and $P(\mathrm{E} \mid {\sim}\mathrm{C_1}, {\sim}\mathrm{C_2}) = 0$.

[8] We thank Ben Rottman for pointing this to us.

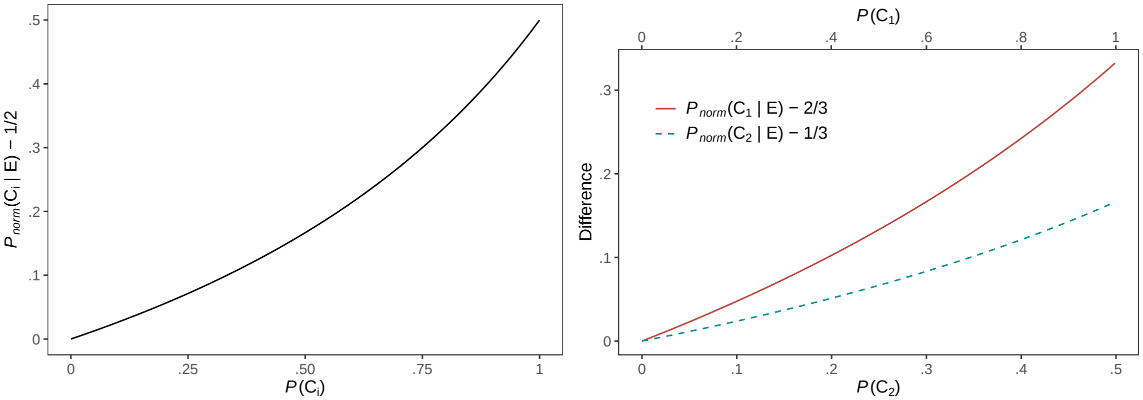

Figure 2: Left: the difference between the normative diagnostic reasoning ($P_{norm}(\mathrm{C_i} \mid \mathrm{E})$) and the constant diagnostic split prediction of $1/2$ in the case of equal priors. Right: the difference between the normative diagnostic reasoning ($P_{norm}(\mathrm{C_1} \mid \mathrm{E})$ and $P_{norm}(\mathrm{C_2} \mid \mathrm{E})$) and the constant diagnostic split predictions of $2/3$ and $1/3$ for $2 : 1$ ratio of the priors. Both figures assume deterministic set-up, i.e., $P(\mathrm{E} \mid \mathrm{C_1}, \mathrm{C_2}) = P(\mathrm{E} \mid \mathrm{C_i}, {\sim}\mathrm{C_j}) = 1$, and $P(\mathrm{E} \mid {\sim}\mathrm{C_1}, {\sim}\mathrm{C_2}) = 0$. We can see that as priors are getting closer to 0 the diagnostic split hypothesis is better approximating the normative diagnostic reasoning.

tic split hypothesis. For instance, Rottman and Hastie (2016) report that the highest point in distributions of participants' diagnostic reasoning responses was at .5 (see Figure 6 in Rottman & Hastie, 2016). This was true for both Experiment 1a where the priors were $P(\mathrm{C}_1) = P(\mathrm{C}_2) = .5$ and Experiment 1b where $P(\mathrm{C}_1) = P(\mathrm{C}_2) = .25$, which suggests the use of the diagnostic split strategy. A recent study by Pilditch, Fenton, and Lagnado (2019) tested people for what they call 'the zero-sum fallacy'. The fallacy stipulates that some people treat evidence as a zero-sum game in which alternative independent hypotheses compete for evidential support and evidential support of one hypothesis means disconfirmation of the other. More specifically, the fallacy is based on the false assumption that the two competing independent hypotheses are mutually exclusive and exhaustive and that evidential support for one hypothesis would entail decrease in the evidential support for the other one. Pilditch et al. (2019) found that when evidence was equally predicted by two competing hypotheses, learning that evidence obtains offers no support for either hypothesis. People displayed this kind of reasoning even after introducing an intervention such as explicitly stating that the hypotheses (causes) are non-exhaustive, and it was shown that the results were not driven by participants' believing that the evidence was non-diagnostic. Although Pilditch et al. (2019) did not provide participants with priors and all data was qualitative, assuming perhaps even natural priors of $P(\mathrm{C}_1) = P(\mathrm{C}_2) = .5$, suggests their findings fit predictions of the diagnostic split hypothesis that $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}) = .5$, since given the priors of .5, E would provide no support for either C_1 or C_2. In addition, a diagnostic split would occur given any priors, as according to zero-sum reasoning, the two causes would be considered mutually exclusive and exhaustive which would imply that $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}) = .5$ for any $P(\mathrm{C}_\mathrm{i})$.

In the present work we directly test the diagnostic split hypothesis. In addition to low and medium priors conditions where we expect to replicate our previous findings (i.e. we expect to find $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}) = .5 \geq P(\mathrm{C}_\mathrm{i})$, for $P(\mathrm{C}_1) = P(\mathrm{C}_2) \leq .5$), in Experiment 1 we also introduced a high priors condition ($P(\mathrm{C_i}) > .5$). In this condition, according to our diagnostic split hypothesis, we expect a significant number of participants to report that the probability of the causes reduces upon learning the effect occurred, compared to their prior probabilities. In other words, we expect to find that a number of participants will erroneously say that $P(\mathrm{C_i} \mid \mathrm{E}) = .5 < P(\mathrm{C_i})$ for $P(\mathrm{C_1}) = P(\mathrm{C_2}) > .5$ even though the causes are maximally strong (i.e. their strengths are 1, see Cheng, 1997).

2.2. Probability interpretations

Another large cluster of data from our previous study, consisted of participants who did not alter the probabilities of causes from the priors after learning the effect occurred or after learning the presence of the effect and the other cause. For these participants, $P(\mathrm{C_i}) = P(\mathrm{C_i} \mid \mathrm{E}) = P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ in both medium and low priors conditions. Through inspection of the data, we ascertained that participants were not merely being inattentive during the task as their completion time suggested they did not rush through the task. Furthermore, they provided explanations about their responses where they usually outlined that since the (prior) probability of one cause happening had been explicitly established, it should not change even in the presence of the effect or of the alternative cause. These considerations led us to hypothesize that participants in this cluster may be interpreting probabilities in a specific way.

In the philosophy of statistics literature, one usually finds that probability interpretations are split into at least two classes: epistemological and objective (Gillies, 2000a, 2000b; Hájek, 2012; Popper, 1959).[9] In epistemological interpretations, probability is related to (the incompleteness of) our knowledge. The most famous interpretation within this class is the subjective probability interpretation, according to which probabilities are

[9] Some authors argue that instead of a strict divide between epistemological and objective probability interpretations, there is a continuum of probability interpretations. See, for instance, Gillies (2000a).

identified as degrees of belief of a particular person, meaning that different individuals can hold different degrees of belief (or different belief strengths) about the same event. On the other hand, objective interpretations view probability as a feature of the material world that is independent of our knowledge or our beliefs. Probabilities, according to this interpretation, can in principle be tested using statistical tests. The frequency interpretation is a well-known objective probability interpretation. Here, probabilities are specified as (limit) frequencies with which an outcome occurs in a sequence of similar events.

A lesser-known probability interpretation is the propensity interpretation (Popper, 1959; Giere, 1973), according to which probabilities are propensities (or tendencies and dispositions) of a particular physical system to produce an outcome (Hájek, 2012). To say that an event X occurs with a probability r, i.e. $P(\mathrm{X}) = r$, is to say that the strength of the propensity of a particular chance set-up to produce outcome X on trial L is r (see Giere, 1973).[10] For example, the statement that the probability of a coin to land Heads equals $\frac{1}{2}$ 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 $\frac{1}{2}$. This propensity is objective, it is part of the physical world, and it does not depend on our 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. Imagine that 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. $P(\mathrm{E} \mid \mathrm{C}_1)$, since whether or not the coin lands Heads or Tails *causally* affects the propensity of the light bulb (i.e. another physical system) to turn on and so

[10] For the purposes of this paper we are confining ourselves to what Gillies (2000b) refers to as 'single-case propensity theories' (see for instance Giere, 1973).

it is perfectly plausible that $P(\mathrm{E} \mid \mathrm{C}_1) \neq P(\mathrm{E})$. So far the propensity interpretation and normative account are in agreement. 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 cannot (backward) causally affect the propensity/the coin bias of Coin 1 to land heads, therefore $P(\mathrm{C}_1 \mid \mathrm{E}) = P(\mathrm{C}_1) = \frac{1}{5}$.[11] In the same vein, additionally learning that Coin 2 landed Heads cannot causally influence how Coin 1 landed and thus cannot not change the propensity of Coin 1 to land Heads, i.e. $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2) = P(\mathrm{C}_1 \mid \mathrm{E}) = P(\mathrm{C}_1) = \frac{1}{5}$. Thus according to the propensity interpretation, observing the effect (or another cause) would not change the propensity of the cause in question to happen. This is stark contrast with the normative account where these three probabilities are in general not equal.

However, like the diagnostic split hypothesis, the propensity interpretation has its normative underpinning in the limit. Figure 3 shows that as the priors converge to 1, the normative diagnostic reasoning estimates approach predictions of the propensity interpretation, i.e. that $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}) - P(\mathrm{C}_\mathrm{i}) = 0$. Furthermore, when the explaining away set-up is deterministic (as in our experiments), then even normatively holds true that $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}, \mathrm{C}_\mathrm{j}) = P(\mathrm{C}_\mathrm{i})$. Thus although the propensity interpretation in general does not accord with the normative account, it can, in some instances, well approximate the normative account. For example, from Figure 3 we can see that the propensity interpretation approximates normative diagnostic reasoning within .1 error when the priors are higher than .63. From Figure 2 on the left we can see that the diagnostic split hypothesis approx-

[11] This intuition has been (formally) outlined in Humphreys (1985), who employs it to argue that propensities are inconsistent with The Kolmogorov Axioms of probability and that, by extension, the propensity interpretation of probability cannot serve as the normative basis. This inconsistency is commonly known as ‘the Humphreys’s paradox’ in the literature.

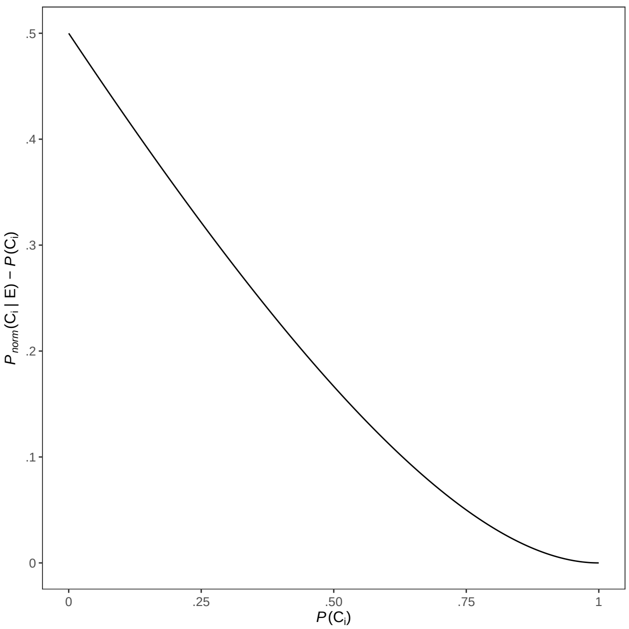

Figure 3: The difference between the normative diagnostic reasoning ($P_{norm}(\mathrm{C_i} \mid \mathrm{E})$) and the prior probability of the causes in the case of equal priors. The figure assumes deterministic set-up, i.e., $P(\mathrm{E} \mid \mathrm{C_1}, \mathrm{C_2}) = P(\mathrm{E} \mid \mathrm{C_i}, {\sim}\mathrm{C_j}) = 1$, and $P(\mathrm{E} \mid {\sim}\mathrm{C_1}, {\sim}\mathrm{C_2}) = 0$. We can see that as priors are getting closer to 1 the propensity interpretation is better approximating the normative diagnostic reasoning.

imate the normative diagnostic reasoning within .1 error when the priors are lower than .33. Thus the propensity interpretation and the diagnostic split hypotheses are complementary to each other: the propensity interpretation well approximates the normative account when the priors are high and the diagnostic split hypothesis does the same when the priors are low. Together, the two are approximating the normative estimates within .1 error for two thirds of all the possible priors. Therefore, even though both are fully opposed to the normative account, together they can reasonably well approximate the normative account.

In this paper we thus hypothesise that the propensity interpretation, which predicts that

$P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) = P(\mathrm{C_i} \mid \mathrm{E}) = P(\mathrm{C_i})$[12], could be partly driving the insufficiency observed in empirical studies of explaining away. The plausibility of this explanation is increased 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). Furthermore, the propensity hypothesis would fit the results reported by Rehder (2014a) where a large proportion (and in most cases the majority) of participants reasoning with a 3-node common effect CBN said that $P(\mathrm{C_1} \mid \mathrm{E})$ is as equally likely as $P(\mathrm{C_1} \mid \mathrm{E}, \mathrm{C_2})$. This was particularly salient in Experiments 1–3 and the deterministic condition of Experiment 4a where no information about the strength of the causal relations was provided to participants, which in turn might have suggested that participants understood the causal relations in these cases to be deterministic: a cause always produces an effect (see Rehder, 2014a).[13]

[12] In general, the propensity interpretation would also predict that $P(\mathrm{C_i}) = P(\mathrm{C_i} \mid \mathrm{C_j}) = P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}) = P(\mathrm{C_i} \mid {\sim}\mathrm{E}) = P(\mathrm{C_i} \mid \mathrm{C_j}, {\sim}\mathrm{E}) = P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}, \mathrm{E}) = P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}, {\sim}\mathrm{E})$. However, given that in this paper we have adopted a deterministic set-up, it is not possible for $P(\mathrm{C_i} \mid {\sim}\mathrm{E})$, $P(\mathrm{C_i} \mid \mathrm{C_j}, {\sim}\mathrm{E})$, and $P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}, {\sim}\mathrm{E})$ to equal $P(\mathrm{C_i})$ since if we learn that evidence does not obtain that means that both causes are false with probability 1. Furthermore, we did not predict that $P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}, \mathrm{E})$ would be accounted by the propensity hypothesis as, in the deterministic set-up, it becomes a simple logic inference (see below) as $P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}, \mathrm{E}) = 1$. Lastly, although the propensity hypothesis predicts that $P(\mathrm{C_i}) = P(\mathrm{C_i} \mid \mathrm{C_j}) = P(\mathrm{C_i} \mid {\sim}\mathrm{C_j})$ we did not focus on these probability estimates when it came to the propensity interpretation (however, see our results sections regarding the independence of causes) as these results are equality predicated by the normative account.

[13] One could argue that even the diagnostic split strategy could be seen as a particular interpretation of probability, namely the classical interpretation according to which the probability of an event is just a fraction of the total number of possibilities in which the event occurs (see Gillies, 2000a; Hájek, 2012). For example, the classical probability of a die landing on an even number is $\frac{3}{6}$. The classical interpretation is thought to

Now, (causal) Bayesian networks (CBNs) usually go hand in hand with the subjective probability interpretation (also referred to as the Bayesian probability interpretation). Pearl (2009, see Section 1.1.2)—as well as Pearl (1988)—is explicit in his adherence to the subjective probability interpretation. Probabilities of nodes in a CBN represent our degrees of belief in events that are causally related and learning that one event happened may affect our degree of belief in some other event (another node in a CBN) happening. On this interpretation, it is perfectly natural to ask both about one's degree of belief the light bulb turned on if the Coin 1 landed Heads as well as one's degree of belief that Coin 1 landed Heads if the light bulb turned on. Moreover, authors empirically testing explaining away, in particular those using CBNs as a benchmark, are explicit about assuming a subjective probability interpretation making comparisons between normative and observed inferences (e.g. Morris & Larrick, 1995; Rehder & Waldmann, 2017). However, people may not always interpret probabilities in a subjective way, which can lead to deviations from the normative account. This sentiment is also expressed by Tversky and Kahneman:

> Decision analysis views subjective probability as a degree of belief, i.e., as a summary of one's state of information about an uncertain event. This concept does not always coincide with the lay interpretation of probability. People

be particularly salient when evidence is symmetrically balanced, which could be expounded as cases where $P(\mathrm{C}_1 \mid \mathrm{E}) = P(\mathrm{C}_2 \mid \mathrm{E}) = \ldots = P(\mathrm{C}_\mathrm{n} \mid \mathrm{E})$. These cases seem to correspond to cases in diagnostic reasoning where participants assign equal probability to each of the possible causes after learning evidence that equally supports each cause. However, as we find that some participants assign unequal probabilities to each cause to reflect unequal priors (Liefgreen et al., 2018), we continue to talk about the diagnostic split strategy rather then the classical interpretations for (i) the classical interpretation has difficulties in handling the cases where the outcomes (possibles) have unequal probabilities, i.e. where outcomes are biased and (ii) the diagnostic split predicts the same probabilities as the classical interpretation when the probabilities of the causes are equal, but also applies to cases where the probabilities of the causes are unequal.

> sometimes think of the probability of an event as a measure of the propensity of some causal process to produce that event, rather than as a summary of their state of belief. The tendency to regard properties as belonging to the external world rather than to our own state of information characterizes much of our perception. We normally regard colors as properties of objects, not of our visual system, and we treat sounds as external rather than internal events. In a similar vein, people commonly interpret the assertion "the probability of heads on the next toss of this coin is 1/2" as a statement about the propensity of the coin to show heads, rather than as a statement about our ignorance regarding the outcome of the next toss. (Tversky & Kahneman, 1977, p. ii)

Testing whether participants' responses on explaining away tasks are partly driven by a particular probability interpretation different from a subjective probability interpretation could then shed light on the findings reported in the extant literature of explaining away.

3. Experiment 1

The aim of this experiment was two-fold: to (i) test people's intuitions in explaining away contexts and (ii) explore if people employ the diagnostic split strategy and/or if they are driven by the propensity interpretation when reasoning in these contexts. In order to do so we used a novel experimental design that not only addressed previously mentioned methodological confounds of previous studies, but additionally allowed us to manipulate two main factors: prior probabilities of causes and the properties of cover stories within which the same common-effect three node structure was embedded.

3.1. Manipulations

3.1.1. Prior probabilities of causes

By manipulating priors of causes we aimed to: (i) vary the amount of normative explaining away (the lower the priors the higher the normative amount of explaining away) and (ii) test the diagnostic split hypothesis. As such, we created three conditions in which the prior probabilities of the causes were either low—$P(\mathrm{C}_1) = P(\mathrm{C}_2) = .2$—medium—$P(\mathrm{C}_1) = P(\mathrm{C}_2) = .5$—or high—$P(\mathrm{C}_1) = P(\mathrm{C}_2) = .7$. In all conditions, the presence of at least one cause entailed the presence of the effect: $P(\mathrm{E} \mid \mathrm{C}_1, \mathrm{C}_2) = P(\mathrm{E} \mid \mathrm{C}_\mathrm{i}, {\sim}\mathrm{C}_\mathrm{j}) = 1$; and absence of both causes entailed absence of the effect: $P(\mathrm{E} \mid {\sim}\mathrm{C}_1, {\sim}\mathrm{C}_2) = 0$. The deterministic relations between the causes and the effect have as a consequence maximal normative explaining away (for a given prior probability) since $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}, \mathrm{C}_\mathrm{j})$ is equal to the prior probability (i.e. to $P(\mathrm{C}_\mathrm{i})$). Additionally, we hoped that these deterministic relations would facilitate people's ability to engage in both diagnostic reasoning and explaining away.

The lower the prior probabilities of causes are, the larger the normative amount of explaining away (see also Rottman & Hastie, 2016). Given the parameters from the previous paragraph, when the priors are low, the probability change from $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E})$ to $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ is .36 and the probability change from $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ to $P(\mathrm{C}_1 \mid \mathrm{E}, {\sim}\mathrm{C}_2)$ is .8, whereas when the priors are medium these changes were .17 and .5 respectively, and only .07 and .3 when the priors are high. Therefore, manipulating priors allowed us to test the prediction that participants would explain away more when reasoning with low priors than when reasoning with both medium and high priors, and that participants reasoning with medium priors would explain away more than those reasoning with high priors.

Additionally, manipulating prior probabilities of causes allowed us to test the diagnostic split hypothesis. We expected a significant number of participants reasoning with low priors to update the probabilities of the two causes to .5 in diagnostic reasoning questions, i.e. in $P(\mathrm{C_i} \mid \mathrm{E})$; for participants reasoning with medium priors we expected them to stay at .5 for both causes in $P(\mathrm{C_i} \mid \mathrm{E})$; and we expected participants reasoning with high priors to lower the probabilities of causes from priors to .5 in $P(\mathrm{C_i} \mid \mathrm{E})$.

3.1.2. Properties of cover stories

In addition to manipulating prior probabilities of causes, we manipulated the properties of the cover stories. In the present study we employed three different cover stories: one involving coin-tossing, one involving balls and containers, and one involving a dinner party. The cover stories were picked such that the propensity interpretation was most accentuated in the coin-tossing cover story, less so in the ball containers one, and least in the dinner party one.

The propensity interpretation itself does not specify which cover stories would lead to more or less acceptance of that interpretation. In devising our cover stories we followed (i) the philosophy of probability literature and (ii) the general idea outlined Section 2.2 on propensity interpretation that propensities are associated with tendencies of a *physical* system that describes a particular chance set-up and that propensities are often tied with causal (or even causal-mechanistic) relationships. This would then imply that we expect to find the propensity interpretation most pronounced in cover stories that include a description of chance set-ups as physical systems with clear causal-mechanistic relations. The cover stories that do not include psychical systems or casual-mechanistic relationships, such as, for instance, cover stories embedded in certain social contexts would render the propensity interpretation less pronounced.

The first cover story where we believed the propensity interpretation would be the highly pronounced included a coin-tossing scenario with the two causes (C_1 and C_2) being represented by two coins (binary variables assuming the value of either Heads or Tails) that are tossed with the same probability p_i for Heads by two coin-tossing mechanisms located in separate rooms. If at least one coin landed Heads,a light bulb (common effect), stored in a different unit and connected to the two coin-tossing mechanisms via electric cables, 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: learning that the light bulb is on/off does not affect the propensity/the disposition for a coin to land Heads. As the questionnaire prompted participants to answer diagnostic reasoning and explaining away questions pertaining to the *coins* (see Section 3.2 below) that are embedded in two physical systems with clear causal-mechanistic relationships to the light bulb we argue that the propensity interpretation will be strongly pronounced in this scenario.

The second cover stories we used included balls and containers where the two causes were represented by two balls (binary variables assuming the value of either copper or rubber) randomly selected from two 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 (the common effect) would turn on. This cover story also included physical systems (mechanisms for random selection of balls from containers) with clear causal-mechanistic relationships (electric circuit) with the common effect (i.e. the light bulb). However, here we follow Giere (1973) in arguing that although the propensity is still present in this cover story, it is at the level of a random sampling mechanism (i.e. the way the balls are selected from the containers), not at the level of balls that are placed onto the electric circuit. The balls are either copper or rubber; they do not have a propensity to be copper or rubber (or if they do it is an extreme propensity of 0 or 1). The random sampling mechanism, on the other hand, does have a propensity p_i to select a copper or a rubber ball from a container. Since, in our study, we prompted participants to answer diagnostic reasoning and explaining away 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 cover story where the propensity was at the level of events in we asked in our questionnaire, namely coins.

Finally, we created a cover story that incorporated a social context namely a dinner party where 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 assuming the value of either 'drinking wine' or 'not drinking wine'). In this cover story, the probability p_i of whether a person brings wine to the party was determined purely by *the subjective estimates* of a host of the party and not by any particular physical system with clear underlying causal-mechanistic relationships to the common cause. For this reason, we argue that in this scenario the propensity interpretation is the least pronounced (if at all).

Given the above rationale, we predicted that the proportion of participants whose reasoning aligns with the propensity interpretation, i.e. who would respond $P(\mathrm{C_i}) = P(\mathrm{C_i} \mid \mathrm{E}) = P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$, 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.

3.2. Methods

3.2.1. Participants and Design

A total of 464 participants ($N_{\mathrm{MALE}} = 181$, $M_{\mathrm{AGE}} = 34.6$ years) 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 10.6 minutes to complete. Eleven participants were excluded as they answered incorrectly to the catch trial, leaving a total of 453 participants in the analyses.

A between-participant design was employed and participants were randomly allocated to one of 3 (Cover story: coins, ball containers, dinner party) × 3 (Priors condition: low, medium or high) = 9 groups ($N_{\text{COINS\_LOW}} = 49$, $N_{\text{COINS\_MED}} = 50$, $N_{\text{COINS\_HIGH}} = 50$, $N_{\text{BALL\_CONTAINERS\_LOW}} = 51$, $N_{\text{BALL\_CONTAINERS\_MED}} = 52$, $N_{\text{BALL\_CONTAINERS\_HIGH}} = 52$, $N_{\text{DINNER\_LOW}} = 50$, $N_{\text{DINNER\_MED}} = 50$, $N_{\text{DINNER\_HIGH}} = 49$).

3.2.2. Materials

Each of the groups was asked to complete an inference questionnaire ($N_{\text{QUESTIONS}} = 12$), comprising of questions regarding priors and (unconditional) independence of causes, as well as reasoning questions relating to diagnostic reasoning and explaining away. For a full list of questions and the inferences these represented see Table 1. For some inferences, such as Diagnostic Reasoning and Explaining away, two questions were asked regarding the same inference, one in qualitative format and one in quantitative format.

As mentioned in Section 3.2.1, participants in each group were required to reason with different cover stories within which we additionally manipulated the priors of causes in the common-effect structure. Three of the groups ($\text{Group}_{\text{COINS\_LOW}}$, $\text{Group}_{\text{COINS\_MED}}$, $\text{Group}_{\text{COINS\_HIGH}}$) reasoned with a coin-tossing cover story in which the two causes (C_1 and C_2) were represented by two simultaneously tossed coins (binary variables assuming the value of either Heads or Tails) in separate rooms and the common effect took the form of a light bulb (LB) in a different unit, that could switch on depending on the outcome of the tosses (if at least one coin landed Heads, the light bulb turns on). An additional three groups ($\text{Group}_{\text{BALL\_CONTAINERS\_LOW}}$, $\text{Group}_{\text{BALL\_CONTAINERS\_MED}}$, $\text{Group}_{\text{BALL\_CONTAINERS\_HIGH}}$) were reasoned with a cover story within which the two causes were represented by two balls (binary variables assuming the value of either copper or rubber) simultaneously drawn from two independent containers and the common effect was again a light bulb in a separate electric circuit, that could switch on depending on the outcome of the draw (if at least one of the balls placed in the circuit is copper, the light bulb turns on). Finally, the remaining three groups (Group$_{\text{DINNER\_LOW}}$, Group$_{\text{DINNER\_MED}}$, Group$_{\text{DINNER\_HIGH}}$) were presented with a cover story in which 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 a dinner party ('Helen' was a binary variable assuming the value of either 'drinking wine' or 'not drinking wine'). For full materials visit Open Science Framework, https://osf.io/aqjkp/.

Table 1: Inference types and questions found in the questionnaire for Experiment 1.

Question Number	Inference Type	Key Inferences	Question Type
1	**Priors**	$P(\mathrm{C}_1)$	Quantitative
2		$P(\mathrm{C}_2)$	Quantitative
3	**Independence**	$P(\mathrm{C}_2 \mid \mathrm{C}_1)$	Qualitative
4		$P(\mathrm{C}_1 \mid {\sim}\mathrm{C}_2)$	Qualitative
5 , 6	**Diagnostic Reasoning**	$P(\mathrm{C}_1 \mid \mathrm{E})$-$R$-$P(\mathrm{C}_1)$	Qual. + Quant.
7 , 8		$P(\mathrm{C}_2 \mid \mathrm{E})$-$R$-$P(\mathrm{C}_2)$	Qual. + Quant.
9 , 10	**Explaining Away**	$P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$-$R$-$P(\mathrm{C}_1 \mid \mathrm{E})$	Qual. + Quant.
11 , 12	**Logic**[14]	$P(\mathrm{C}_1 \mid \mathrm{E}, {\sim}\mathrm{C}_2)$-$R$-$P(\mathrm{C}_1 \mid \mathrm{E})$	Qual. + Quant.

Note: -*R*- stands for 'in relation to'.

3.2.3. Procedure

Participants in each of the nine 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. In the coins and the dinner party cover stories the priors were presented the form of a percentage, whereas in the ball containers cover story they were presented as a fraction/ratio (e.g. of the 10 balls, there are 2 copper balls and 8 rubber balls in each urn).[15] The priors in cases of the coins and the dinner party cover stories were given only in a textual form. The priors in the ball container cover story (i.e. the number of ball of each type) and the causal relations in all cover stories were given to participants in both textual form and in visual form (graphical representation). In order to ensure participants understood the structure, they 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 see Table 1). The questionnaire required participants to *sequentially* answer questions: firstly regarding priors of causes, secondly independence of causes, thirdly diagnostic reasoning about each cause, and finally regarding explaining away. The graphical and textual details of the cover story were present on the same page as the relevant inference questions so participants could access these details at any point.

Questions marked as quantitative in Table 1 required participants to provide numerical estimates on a slider with a scale ranging from 0% to 100%. Questions marked as

[14] We have labeled questions 11 and 12 as ‘logic’ questions, since our set-up was deterministic and learning that one cause did not happen, whilst knowing that the effect happened, entailed (*by logic*) that the other cause must have happened, i.e. $P(\mathrm{C}_1 \mid \mathrm{E}, {\sim}\mathrm{C}_2) = 1$.

[15] Although the way priors were conveyed depended on a cover story, in all cover stories they were elicited in the same manner, i.e. as a percentage on a scale from 0% to 100%.

qualitative, required participants to select one of three options: the probability increases, decreases, or stays the same when asked about e.g. $P(\mathrm{C}_2 \mid \mathrm{C}_1)$ given no knowledge of whether E is present or not. To investigate participants' diagnostic and explaining away reasoning we employed both qualitative and quantitative question formats. For example, participants in the coin tossing cover story, after finding out that the light bulb is on, were asked both a *qualitative* diagnostic reasoning question (e.g. Q5): "Does the probability that **Coin 1** landed Heads **change** (compared to Q1, where you said: X%) after you find out that the light bulb turned on?" as well as a *quantitative* one: "What do you now think is the probability that **Coin 1** landed Heads?". This approach enabled us to capture the relational nature of explaining away, as well as 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. All evidence (i.e. new states of cause or effect variables) was provided to participants both textually (e.g. in groups reasoning with a coin tossing cover story: "You walk into Unit 3 and see that the light bulb is on") as well as visually (as an updated graphical representation of the model).

3.3. Results

Participants' answers to all qualitative questions in the inference questionnaire are represented in Figure 4 and their responses to all quantitative questions are in Figure 5.

3.3.1. Overall performance

To test for a main effect of cover story and/or priors on participants' judgment accuracy we initially coded all participants' answers as correct (1) or incorrect (0). For all quantitative estimates, an answer was considered correct if it fell within ± .02 of the normative probability estimate. This allowed us to have a comparative measure of participants' accuracy for both qualitative and quantitative types of inferences. Subsequently, if an inference

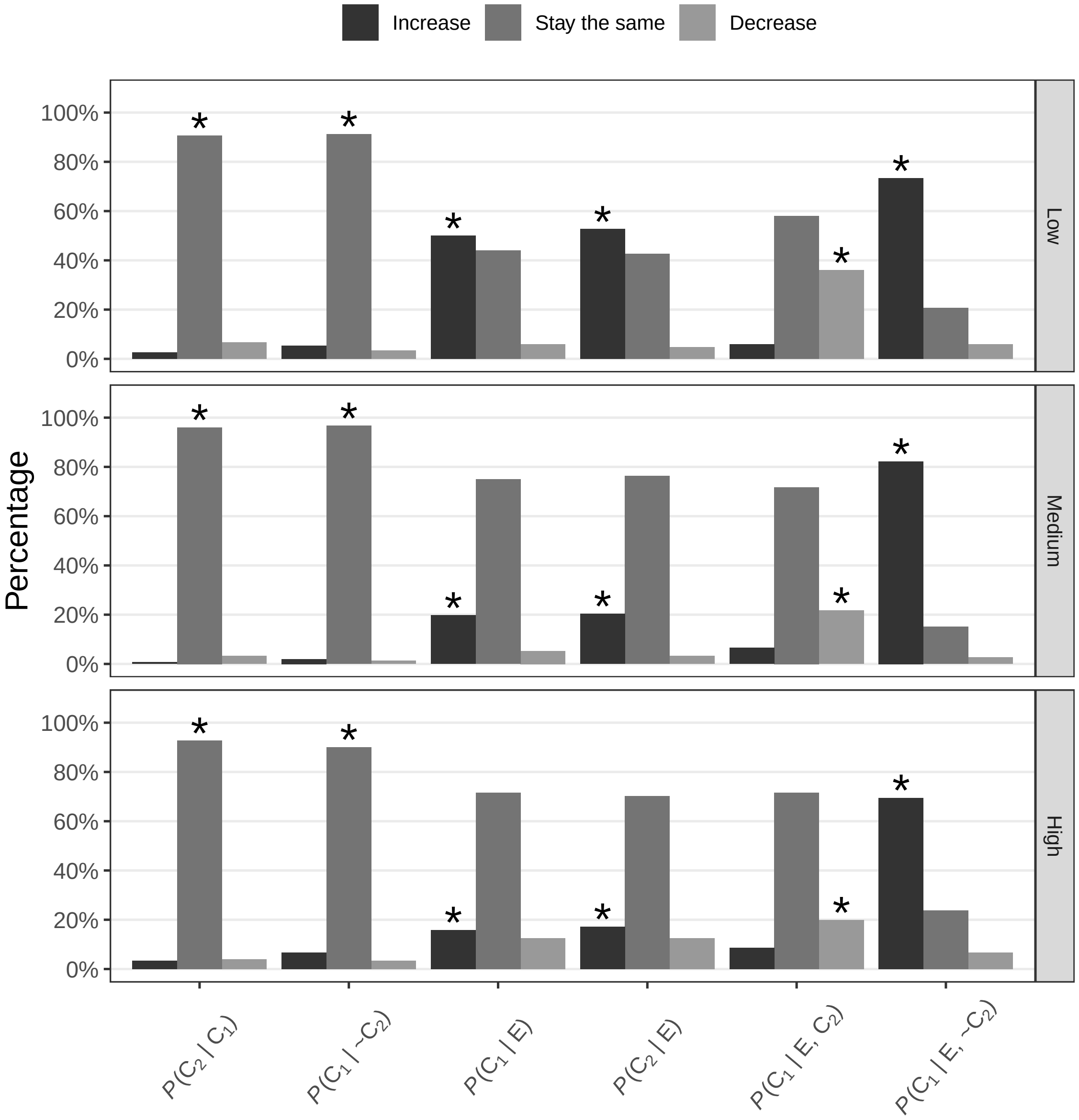

Figure 4: Distribution of participants' responses to qualitative questions in Experiment 1. Asterisks above the bars indicate normative answers.

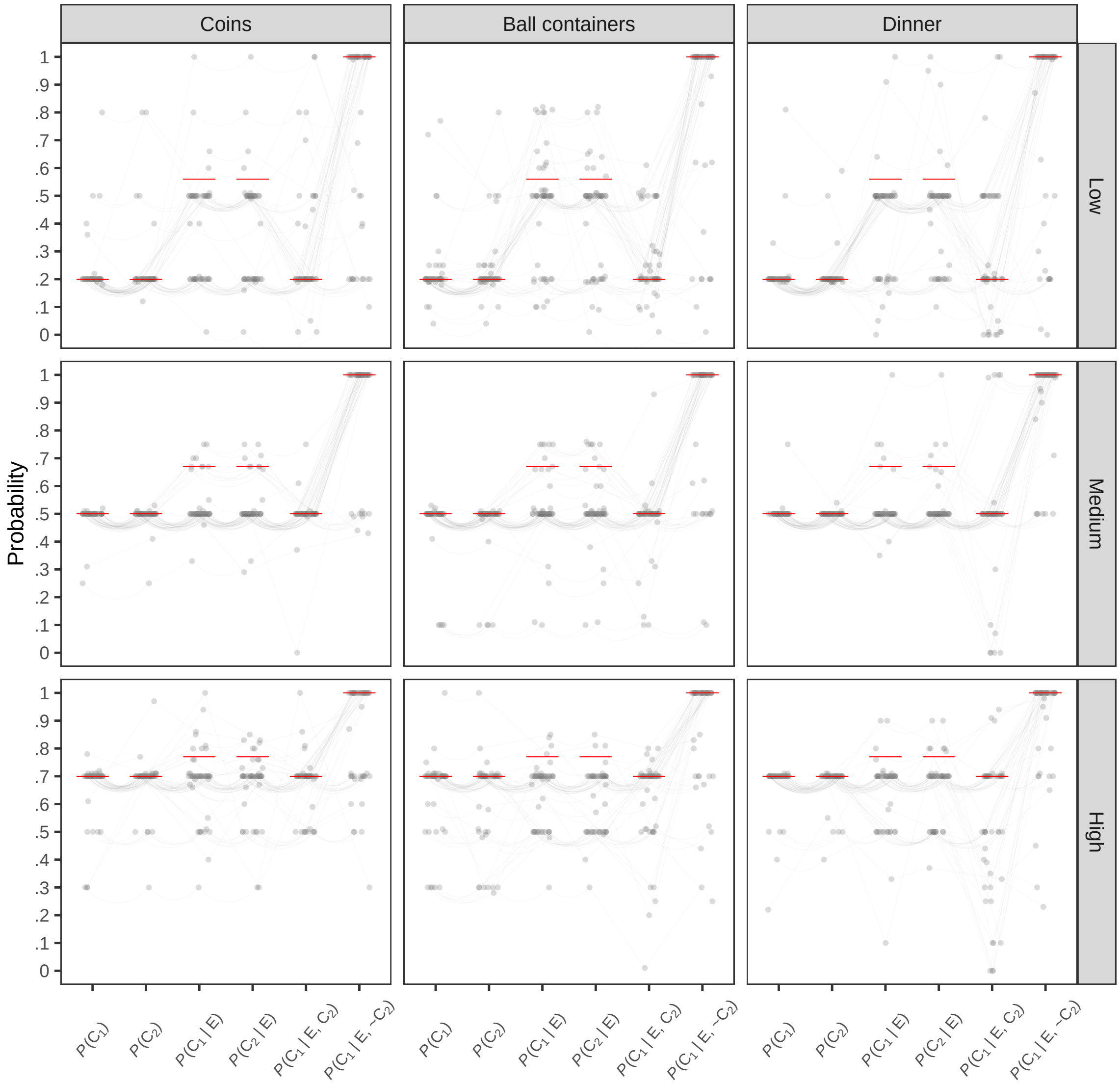

Figure 5: Participants' responses to quantitative questions in Experiment 1. Red horizontal lines are correct (normative) answers. Gray lines between data points depict how participants changed their probability estimates from one questions to another, with curved lines indicating that a participant did not change (within ± .02) their probability estimate.

judgment had a symmetrical pair, i.e. if both inference judgements were of the same inference type (such as inferences regarding priors, independence, qualitative, and quantitative diagnostic reasoning, see Table 1) we combined each participant's coded response to both questions into a single coded response: if a participant answered both questions correctly, the response was coded as 1; otherwise 0. This left us with eight coded question-types regarding: priors, independence, qualitative diagnostic reasoning, quantitative diagnostic reasoning, qualitative explaining away, quantitative explaining away, qualitative logic, and quantitative logic. For descriptive statistics of participant accuracy within each group see Table 2 below.

Table 2: Descriptive statistics of participants' overall performance per group in Experiment 1.

Group	Proportion Correct Answers	95% CI
Coins Low	.55	[.48, .62]
Coins Med	.61	[.56, .66]
Coins High	.48	[.41, .54]
Balls Containers Low	.55	[.49, .62]
Balls Container Med	.58	[.53, .64]
Balls Containers High	.49	[.44, .54]
Dinner Party Low	.58	[.52, .63]
Dinner Party Med	.59	[.55, .62]
Dinner Party High	.51	[.47, .55]

To test the effect of Cover story and Priors on participants' overall performance (in the coded form) on the eight question-types, we built a generalized linear mixed effects model with a binomial link function using the lme4 package (Bates, Mächler, Bolker, & Walker, 2014). The model had two fixed effects, Cover story and Priors, with a random intercept for each participant (there was no random slope for participant since Cover story and Priors vary between participants). We found a main effect of Priors, $z = -3$, $p = .003$ and no main effect of Cover story, $z = 0.56$, $p = .58$. We also found no interaction between Cover story and Priors, $z = 0.12$, $p = .9$. Including the predictors (Cover story and Priors) in the model did improve model fit ($\chi^2(3) = 9.33$, $p = .025$) compared to just having an intercept as a predictor.

Given that in the above analyses we found no main effect of Cover story on accuracy nor an interaction between Cover story and Priors, we collapsed data across cover stories to perform the subsequent analyses regarding participants' performance on explaining away. Therefore, we now compare across three groups : a low priors group (Group$_{\text{LOW}}$, $N = 150$), a medium priors group (Group$_{\text{MEDIUM}}$, $N = 152$), and a high priors group (Group$_{\text{HIGH}}$, $N = 151$).

3.3.2. Acceptance of priors

A Chi-Square test of independence illustrated a significant difference in the proportion of participants who accurately stated the priors of *both* causes between the three groups, $\chi^2(2) = 12.9$, $p = .002$. Post-hoc pairwise comparisons using Benjamini and Hochberg's (1995) false discovery rate (FDR) procedure with $q^* = 0.05$[16] indicated a significant difference between Group$_{\text{MEDIUM}}$ (92.8%) and Group$_{\text{HIGH}}$ (78.1%), corrected $p = .002$. No significant difference was found between Group$_{\text{LOW}}$ (84.7%) and either

[16] The same applied to all other pairwise comparisons in the paper.

Group$_{\mathrm{MED}}$, corrected p = .062, and Group$_{\mathrm{HIGH}}$, corrected p = .192.

In addition, for each participant we computed an absolute difference from the stated priors. Since the data are quite clearly non-normally distributed (Figure 5) we adopted non-parametric tests. A Kruskal–Wallis test illustrated a significant effect of Priors on the absolute value differences, $H(2) = 31.9$, $p < .001$. Pairwise comparisons of the mean ranks between groups showed a significant difference between Group$_{\mathrm{MEDIUM}}$ and both Group$_{\mathrm{HIGH}}$ ($difference = 83.2$, $critical\, difference = 50.9$[17]) and Group$_{\mathrm{LOW}}$ ($difference$ = 56.1, $critical\, difference = 51$); the difference between Group$_{\mathrm{HIGH}}$ and Group$_{\mathrm{LOW}}$ was not significant ($difference = 27$, $critical\, difference = 51.1$). Though the difference between the above groups was significant, the high proportion of participants who stated the correct priors for both causes and the low absolute differences from the stated priors within each group indicate that overall participants accepted priors of causes given to them, across all conditions (see also the distributions of participants responses for $P(\mathrm{C}_1)$ and $P(\mathrm{C}_2)$ in Figure 5).

3.3.3. Independence of causes

For a breakdown of the frequency of participants' choices on independence questions see Figure 4. Within each group we obtained the percentage of people who correctly answered *both* questions regarding the independence of causes (Q3 and Q4 in Table 1). Within Group$_{\mathrm{LOW}}$ this was 88.7%, within Group$_{\mathrm{MEDIUM}}$ this was 95.4% and within Group$_{\mathrm{HIGH}}$ this was 88.1%. These high percentages demonstrate that the vast majority of participants did not violate the assumption of the independence of causes (before learning the evidence) in any group.

[17] Throughout the paper, the critical difference at $\alpha = .05$ was corrected for the number of tests.

3.3.4. Diagnostic reasoning

Independent analyses were conducted on qualitative and quantitative diagnostic reasoning questions (Qs 5–8 in Table 1).

Qualitative. A Chi-Square test of independence illustrated a significant difference in the proportion of participants who accurately answered *both* qualitative questions relating to diagnostic reasoning between the three groups, $\chi^2(2) = 52.27$, $p < .001$. Post-hoc pairwise comparisons using the FDR procedure illustrated a significant difference between $\text{Group}_{\text{LOW}}$ (45.3%) and both $\text{Group}_{\text{MEDIUM}}$ (17.1%), corrected $p < .001$ and $\text{Group}_{\text{HIGH}}$ (11.9%), corrected $p < .001$. No significant difference was found between $\text{Group}_{\text{MEDIUM}}$ and $\text{Group}_{\text{HIGH}}$, corrected $p = .26$. As can be seen from Figure 4 almost half of the participants in $\text{Group}_{\text{LOW}}$ indicated the change of probability in the correct direction, which significantly differed from the percentage of participants in $\text{Group}_{\text{MEDIUM}}$ and $\text{Group}_{\text{HIGH}}$. This is an interesting finding as it seems to suggest that a larger normative quantitative difference between the two probabilities corresponds to a larger proportion of participants following the normative qualitative direction. Here we have that the largest probability increase was in the low priors condition: $P(\text{C}_\text{i} \mid \text{E}) - P(\text{C}_\text{i}) = .36$, followed by the medium priors condition where the increase was .17 and the high priors condition where it was only .07. The size of these quantitative normative quantitative difference between the two probabilities directly corresponded to size of the proportions of participants who answered the qualitative questions in accordance with the normative model.

Quantitative. Fischer's exact test of independence illustrated a significant difference in the proportion of participants who correctly answered *both* quantitative diagnostic reasoning questions across the three groups, $p = .002$. Post-hoc pairwise comparisons using the FDR procedure illustrated a significant difference between $\text{Group}_{\text{LOW}}$ (0%) and $\text{Group}_{\text{MEDIUM}}$ (6.6%) , corrected $p = .005$. No significant difference was found between $\text{Group}_{\text{HIGH}}$

(2.6%) and both $\text{Group}_{\text{LOW}}$, corrected $p = .17$, and $\text{Group}_{\text{MEDIUM}}$, corrected $p = .17$. The low percentages suggest that all groups performed poorly compared to the normative model (see also the distributions of responses for $P(\text{C}_1 \mid \text{E})$ and $P(\text{C}_2 \mid \text{E})$ in Figure 5).

To gauge how much participants deviated from the normative estimates, we computed a sample standard deviation from the normative response (s_{norm}) for each group. On $P(\text{C}_1 \mid \text{E})$ question, for $\text{Group}_{\text{LOW}}$, $s_{norm} = 24.5$, 95% CI [22.1, 27][18]; for $\text{Group}_{\text{MEDIUM}}$, $s_{norm} = 17.9$, 95% CI [16.5, 20.3]; for $\text{Group}_{\text{HIGH}}$, $s_{norm} = 17.1$, 95% CI [15.1, 20.1]. On $P(\text{C}_2 \mid \text{E})$ question, for $\text{Group}_{\text{LOW}}$, $s_{norm} = 24.2$, 95% CI [21.9, 26.5]; for $\text{Group}_{\text{MEDIUM}}$, $s_{norm} = 17.9$, 95% CI [16.4, 20.3]; for $\text{Group}_{\text{HIGH}}$, $s_{norm} = 16.9$, 95% CI [15, 19.1]. This suggests that $\text{Group}_{\text{LOW}}$ most deviated from the normative answers compared to the other two groups. This is expected from the normative perspective since the normative amount of diagnostic reasoning (the difference between $P(\text{C}_\text{i})$ and $P(\text{C}_\text{i} \mid \text{E})$) is the highest in $\text{Group}_{\text{LOW}}$.

We also explored the amount and direction of change in participants' probabilistic estimates from their given priors to their estimates after learning about the effect. As such we conducted the Wilcoxon signed-rank test on the difference between participants' estimates on each prior question and the related diagnostic reasoning question (i.e. between $P(\text{C}_1)$ and $P(\text{C}_1 \mid \text{E})$ and between $P(\text{C}_2)$ and $P(\text{C}_2 \mid \text{E})$). When comparing these differences with the normative differences, the null hypotheses of all Wilcoxon signed-rank tests was that the difference between participants' estimates equals to the corresponding normative difference. Table 3 shows the normative differences, the empirical differences of medians, and p-values of Wilcoxon signed-rank tests.

As can be seen from the table, participants heavily under-adjusted their probability esti-

[18] The 95% confidence intervals were calculated using the BCa nonparametric bootstrap confidence interval method (with 10^6 bootstrap replicates) as recommend by Meeker, Hahn, and Escobar (2017).

Table 3: Quantitative differences in diagnostic reasoning inferences per group in Experiment 1.

Inferences	Normative difference	Empirical difference of medians	p-value
Group$_{\text{LOW}}$			
$P(\text{C}_1 \mid \text{E}) - P(\text{C}_1)$	.36	.3	< .001
$P(\text{C}_2 \mid \text{E}) - P(\text{C}_2)$	.36	.3	< .001
Group$_{\text{MEDIUM}}$			
$P(\text{C}_1 \mid \text{E}) - P(\text{C}_1)$	.17	0	< .001
$P(\text{C}_2 \mid \text{E}) - P(\text{C}_2)$	.17	0	< .001
Group$_{\text{HIGH}}$			
$P(\text{C}_1 \mid \text{E}) - P(\text{C}_1)$	.07	0	< .001
$P(\text{C}_2 \mid \text{E}) - P(\text{C}_2)$	.07	0	< .001

mates since the null hypothesis that the normative difference is equal to the empirical difference is strongly rejected in all cases. Furthermore, only in Group$_{\text{LOW}}$ did the empirical difference go in the normative direction. In both Group$_{\text{MEDIUM}}$ and Group$_{\text{HIGH}}$ the empirical differences of medians was 0 suggesting that in these groups participants' quantitative diagnostic reasoning estimates did not significantly differ from their priors estimates.

3.3.5. Direct explaining away

Independent analyses were conducted on qualitative and quantitative questions regarding direct explaining away (Q9 and Q10 in Table 1).

Qualitative. For a breakdown of the frequency of participants' choices on the qualitative direct explaining away question see Figure 4. A Chi-Square test of independence illustrated a significant difference in the proportion of participants who accurately answered the qualitative question relating to explaining away between the three groups, $\chi^2(2) = 12.25$, $p = .002$. Similarly to the results regarding diagnostic reasoning (Section 3.3.4), post-hoc pairwise comparisons using the FDR procedure illustrated a significant difference between Group$_{\text{LOW}}$ (36%) and both Group$_{\text{MEDIUM}}$ (21.7%), corrected $p = .013$ and Group$_{\text{HIGH}}$ (19.9%), corrected $p = .008$. No significant difference was found between Group$_{\text{MEDIUM}}$ and Group$_{\text{HIGH}}$, corrected $p = .8$. This suggests that participants in Group$_{\text{LOW}}$ performed significantly better than participants in Group$_{\text{MEDIUM}}$ and participants in Group$_{\text{HIGH}}$. Similarly to qualitative diagnostic reasoning, this was congruent with the the size of the normative explaining found in the respective Priors conditions. Overall, however, the low percentage of correct responses across groups suggest poor performance in this category.

Quantitative. A Chi-Square test of independence illustrated a significant difference in the proportion of participants who accurately answered the quantitative question regarding direct explaining away between the three groups, $\chi^2(2) = 34.74, p < .001$. Post-hoc pairwise comparisons using the FDR procedure illustrated a significant difference between Group$_{\text{MEDIUM}}$ (82.9%), and both Group$_{\text{LOW}}$ (52.7%), corrected $p < .001$ and Group$_{\text{HIGH}}$ (57.6%), corrected $p < .001$. No significant difference was found between Group$_{\text{LOW}}$ and Group$_{\text{HIGH}}$, corrected $p = 0.46$. This suggests that in each group over half of the participants correctly answered the direct explaining away question.

For each group we also computed a sample standard deviation from the normative response (s_{norm}). For Group$_{\text{LOW}}$, $s_{norm} = 22.4$, 95% CI [18.7, 27.2]; for Group$_{\text{MEDIUM}}$, $s_{norm} = 15.2$, 95% CI [12, 18.9]; for Group$_{\text{HIGH}}$, $s_{norm} = 20.8$, 95% CI [17.3, 25]. This suggests that Group$_{\text{MEDIUM}}$ least deviated from the normative answers compared to the other two groups. The relatively high percentages of correct answers and a relatively low deviation from the normative answers may suggest good performance on quantitative direct explaining away. Although this may appear as being at odds with our finding of overall poor performance on qualitative direct explaining away, a quick look at Figure 5 reveals that a large number of participants repeated the priors in $P(\mathrm{C}_1 \mid \mathrm{E})$, $P(\mathrm{C}_1 \mid \mathrm{E})$, and $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ (this is discussed in Section 3.3.9 below). Since in our study $P(\mathrm{C}_1) = P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ and a large proportion of participants did accept the priors (see Section 3.3.2), this suggests that a large proportion did correctly answer the quantitative direct explaining question. This result highlights the importance of also including qualitative relational questions in such contexts.

3.3.6. Logic

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

Qualitative. A Chi-Square test of independence illustrated a significant difference in the proportion of participants who accurately answered the qualitative question relating to explaining away between the three groups, $\chi^2(2) = 6.88$, $p = .032$. Post-hoc pairwise comparisons using FDR procedure illustrated a significant difference between Group$_{\mathrm{MEDIUM}}$ (82.2%) and Group$_{\mathrm{HIGH}}$ (69.5%), corrected $p = .043$. No significant difference was found between Group$_{\mathrm{LOW}}$ (73.3%) and both Group$_{\mathrm{MEDIUM}}$, corrected $p = .127$, and Group$_{\mathrm{HIGH}}$, corrected $p = .548$. As can be seen from Figure 4, the majority of participants did, however, correctly report the direction of the probability change.

Quantitative. A Chi-Square test of independence illustrated no significant difference in the proportion of participants who accurately answered the quantitative question relating to explaining away between the three groups, $\chi^2(2) = 4.26$, $p = .119$. The proportions were: Group$_{\text{LOW}}$, 68.7%; Group$_{\text{MEDIUM}}$, 77%; and Group$_{\text{HIGH}}$, 66.9%. The high percentages suggest that in each group a majority of the participants correctly answered the logic question.

Overall these findings illustrate that across conditions a high percentage of participants was able to correctly answer both quantitative and qualitative logic questions, suggesting they largely understood the (deterministic) relations between variables in the 3-node structure.

3.3.7. 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. Independent analyses were conducted on qualitative and quantitative relational explaining away questions.

Qualitative. To explore participants' qualitative relational explaining away, we conducted the analysis on questions relating to direct explaining away and logic (Q9 and Q11 in Table 1).[19] A Chi-Square test of independence illustrated a significant difference in the proportion of participants who accurately answered both qualitative questions relating to explaining away concept between the three groups, $\chi^2(2) = 12.8$, $p = .002$. Post-hoc pairwise comparisons using the FDR procedure illustrated a significant difference between Group$_{\text{LOW}}$ (32.7%) and Group$_{\text{HIGH}}$ (15.9%), corrected $p = .003$ and between Group$_{\text{LOW}}$ and Group$_{\text{MEDIUM}}$ (20.4%), corrected $p = .033$. No significant difference was found between Group$_{\text{MEDIUM}}$ and Group$_{\text{HIGH}}$, corrected $p = .386$. Similarly to the qualitative diag-

[19] We did not include the two qualitative diagnostic reasoning questions here since these two questions are about the relationship between the priors and diagnostic reasoning. Our aim was to analyze participants understanding of the inequalities in (2) which are about the relations between diagnostic reasoning and direct explaining away (Q9) and between direct explaining away and 'logic' (Q11).

nostic reasoning and the qualitative direct explaining away results, these proportions seem to correspond to the size of the normative relational explaining away in respective Priors conditions. The percentages, however, are again low suggesting poor overall performance.

Quantitative. In regards to the quantitative relational explaining away, the questions we included in the analyses were those relating to the updating of C_1, namely, $P(\mathrm{C}_1 \mid \mathrm{E})$, $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$, and $P(\mathrm{C}_1 \mid \mathrm{E}, {\sim}\mathrm{C}_2)$. These are Q6, Q10 and Q12 in Table 1.

A Friedman's ANOVA was carried out on participants' estimates of the quantitative relational explaining away questions, within each of the groups (see Figure 6). Results illustrated a significant difference between these estimates within $\mathrm{Group}_{\mathrm{LOW}}$, $\chi^2(2) = 155.9$, $p < .001$, within $\mathrm{Group}_{\mathrm{MEDIUM}}$, $\chi^2(2) = 190.9$, $p < .001$ and within $\mathrm{Group}_{\mathrm{HIGH}}$, $\chi^2(2) = 157.2$, $p < .001$.

Wilcoxon signed-rank tests were carried out to compare participants' estimates with normative ones (see Table 4 below). In each of the tests, the null hypothesis was that the empirical difference between the pairs of inferences of interest would equal the corresponding normative difference . As can be seen from the table, participants mostly under-adjusted their probability estimates since the null hypothesis that the normative difference is equal to the empirical difference is strongly rejected in most cases except in $\mathrm{Group}_{\mathrm{HIGH}}$ between $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ and $P(\mathrm{C}_1 \mid \mathrm{E}, {\sim}\mathrm{C}_2)$ where participants appear to have sufficiently shifted their estimates. The participants in $\mathrm{Group}_{\mathrm{LOW}}$ and $\mathrm{Group}_{\mathrm{MEDIUM}}$ have thus under-adjusted their estimates despite the difference in medians between $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ and $P(\mathrm{C}_1 \mid \mathrm{E}, {\sim}\mathrm{C}_2)$ being equal to the normative difference for these groups.

3.3.8. Diagnostic split

To test the diagnostic split hypothesis we included in our analysis only participants who reported the correct priors and then calculated the proportion of these participants who reported .5 (± .02) as their estimate for *both* $P(\mathrm{C}_1 \mid \mathrm{E})$ and $P(\mathrm{C}_2 \mid \mathrm{E})$. These were:

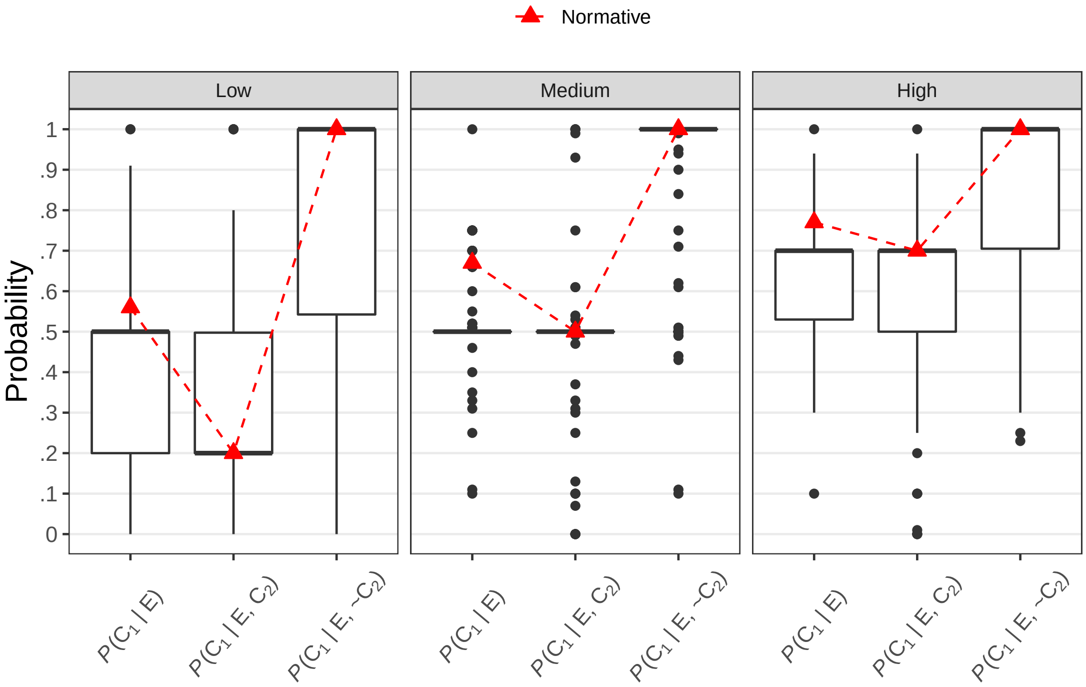

Figure 6: Box plots of participants' quantitative relational explaining away responses in three groups along with the normative estimates in Experiment 1.

Table 4: Within-group explaining away in Experiment 1.

Inferences	Normative difference	Empirical difference of medians	p-value
$\mathit{Group}_{\text{LOW}}$			
A – B	.36	.3	$< .001$
C – B	.8	.8	$< .001$
$\mathit{Group}_{\text{MED}}$			
A – B	.17	0	$< .001$
C – B	.5	.5	$< .001$
$\mathit{Group}_{\text{HIGH}}$			
A – B	.07	0	$< .001$
C – B	.3	.3	.067

Note: A := $P(\text{C}_1 \mid \text{E})$, B := $P(\text{C}_1 \mid \text{E}, \text{C}_2)$, C := $P(\text{C}_1 \mid \text{E}, {\sim}\text{C}_2)$.

50.4% in Group$_{\text{LOW}}$, 78.7% in Group$_{\text{MED}}$ and 13.6% in Group$_{\text{HIGH}}$. A Chi-Square test of independence illustrated that these proportions significantly differed from each other, $\chi^2(2) = 109.2$, $p < .001$. All post-hoc pairwise comparisons using the FDR procedure were significant with corrected $p < .001$. These proportions suggest that a large proportion of participants who correctly answered the priors questions provided estimates predicted by the diagnostic split hypothesis. Note that both the diagnostic split hypothesis and the propensity hypothesis make exactly the same prediction in the medium priors condition, namely stay at the prior of .5. Therefore, the higher proportion observed in the Group$_{\text{MED}}$ is expected as the .5 response is predicted by both hypotheses. The relatively low proportion of participants observed in Group$_{\text{HIGH}}$ suggests that people are unwilling to reduce the probability to .5 in diagnostic reasoning from the high prior of .7. Overall then, these results partly support the diagnostic split hypothesis.

At the outset of the paper, we predicted that the diagnostic split hypothesis would be able to account for a significant amount of failures in (quantitative) diagnostic reasoning and (quantitative) relational explaining away. To explore how much of these failures can be explained by the diagnostic split hypothesis we built simple cross-tabulations. We selected only participants who correctly answered the both priors questions and collapsed the data across all the condition. We then cross-tabulated participants' responses as in line ('yes') or not in line ('no') with the diagnostic split hypothesis and correct ('yes') or incorrect ('no') quantitative diagnostic reasoning as well as correct ('yes') or incorrect ('no') quantitative relational explaining away (see Table 5) (these tables also included responses that were in line ('yes') or not in line ('no') with the propensity interpretation since this was relevant for the section below).[20] First, notice that the cross-tabulations in for both diagnostic reasoning and explaining way look very similar suggesting that participants who correctly answer the quantitative diagnostic reasoning questions went on to also correctly answer questions related to the quantitative direct explaining and the quantitative logic question. However, as only 13 participants correctly answered the quantitative diagnostic reasoning question this applied to only about 3% of the data. Second, from the table we

[20] We have not included the diagnostic split hypothesis in cross-tabulations that included *qualitative* diagnostic reasoning and *qualitative* relational explaining away, as we did with the propensity hypothesis, since (i) the propensity hypothesis has a very specific quantitative prediction that does not dependent on the qualitative directional of update from the priors and (ii) the diagnostic split hypothesis would have the same qualitative prediction as the normative account in the low priors conditions (i.e. the probability should increase) and in order not to conflate these two we have not included the diagnostic split hypothesis in cross-tabulations on the qualitative results.

can see that the diagnostic split hypotheses accounted for about 51% violations in quantitative diagnostic reasoning and in quantitative relational explaining away. This finding suggests that the diagnostic split reasoning played a significant part in violations of both the quantitative diagnostic reasoning and quantitative relational explaining away.

Table 5: A cross-tabulation for correct/incorrect (yes/no) quantitative diagnostic reasoning and quantitative relation explaining away as well as for in line/not in line with (yes/no) the diagnostic split hypothesis and the (quantitative) propensity hypothesis predictions in Experiment 1.

		Diagnostic reasoning		Explaining away	
		Quantitative		Quantitative relational	
		Yes	*No*	*Yes*	*No*
Diagnostic split	Propensity interpretation (quantitative)				
Yes	*Yes*	0	101	0	101
Yes	*No*	0	90	0	90
No	*Yes*	0	99	0	99
No	*No*	13	83	12	84

3.3.9. Propensity interpretation

In order to test our propensity hypothesis, we calculated the proportion of people who did not update in the face of learning evidence and learning the other cause occurred.

Qualitative. We calculated the proportions of participants who, having stated the correct priors, selected ‘stay the same’ as an answer to both *qualitative* diagnostic reasoning questions (Q5 and Q7) as well as the *qualitative* direct explaining away question (Q9). Across each cover story these percentages were: 63.8% for Group$_{\text{COINS}}$, 53.8% for Group$_{\text{BALL\_CONTAINERS}}$, and 46.8% for Group$_{\text{DINNER}}$. A Chi-Square test of independence found a significant difference between these proportions, $\chi^2(2) = 7.96$, $p = .019$. Post-hoc pairwise comparisons using the FDR procedure showed the only difference to be between Group$_{\text{COINS}}$ and Group$_{\text{DINNER}}$, corrected $p = .021$. No significant difference was found between Group$_{\text{COINS}}$ and Group$_{\text{BALL\_CONTAINERS}}$, corrected $p = .213$, or between Group$_{\text{BALL\_CONTAINERS}}$ and Group$_{\text{DINNER}}$, corrected $p = .316$.

Quantitative. Out of the participants who correctly stated the priors, we calculated the proportions of those who provided the priors as their estimate to $P(\mathrm{C}_1 \mid \mathrm{E})$, $P(\mathrm{C}_2 \mid \mathrm{E})$, and $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ (i.e. Q6, Q8, and Q10). Collapsing across the priors conditions, the percentages were: 60.8% for Group$_{\text{COINS}}$, 50% for Group$_{\text{BALL\_CONTAINERS}}$ and 44.6% for Group$_{\text{DINNER}}$. Chi-Square test of independence illustrated that these proportions significantly differed from each other, $\chi^2(2) = 7.2$, $p = .028$. Post-hoc pairwise comparisons using the FDR procedure showed the only significant difference to be between Group$_{\text{COINS}}$ and Group$_{\text{DINNER}}$, corrected $p = .034$. No significant difference was found between Group$_{\text{COINS}}$ and Group$_{\text{BALL\_CONTAINERS}}$, corrected $p = .198$, or between Group$_{\text{BALL\_CONTAINERS}}$ and Group$_{\text{DINNER}}$, corrected $p = .421$.

The results from the qualitative and quantitative participants' responses fit the propensity hypothesis prediction: we found that significantly more participants stay at the priors in the Coins cover story where we expected the propensity hypothesis to be the most pronounced compared to the Dinner cover story, with the Ball containers cover story falling in between.

Furthermore, from Table 5 we can see that the propensity hypothesis accounted for about 53% of violations in both the quantitative diagnostic reasoning and quantitative relational explaining away. We also cross-tabulated participants' answers as (not) in line with the propensity hypothesis and (in)correct qualitative direct and relational explaining away and (in)correct quantitative direct explaining away. Table 6 shows that the propensity hypothesis accounted for about 73% of violations in qualitative diagnostic reasoning, about 74% of violations in qualitative direct explaining away, and about 71% of violations in qualitative relational explaining away. The high percentages suggest that the propensity hypothesis was driving the majority of violations in all these inferences. Table 7 further elucidates the point from Section 3.3.5 where we found that an unexpectedly large proportion of participants correctly answered the quantitative direct explaining away question. Here we see that about 70% of these 'correct' responses were in fact responses given in line with the propensity hypothesis where participants repeated the priors when answering the quantitative direct explaining away question.

Table 6: A cross-tabulation for correct/incorrect (yes/no) qualitative diagnostic reasoning and both direct and relational qualitative explaining away as well as for in line/not in line with (yes/no) the (qualitative) propensity hypothesis predictions in Experiment 1.

	Diagnostic reasoning		Qualitative explaining away			
	Qualitative		Direct		Relational	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*
Propensity interpretation (qualitative)						
Yes	0	211	0	211	0	211
No	95	80	100	75	89	86

Table 7: A cross-tabulation for correct/incorrect (yes/no) quantitative direct explaining away as well as for in line/not in line with (yes/no) the (quantitative) propensity hypothesis predictions in Experiment 1.

	Quantitative direct explaining away	
	Yes	*No*
Propensity interpretation (quantitative)		
Yes	200	0
No	86	100

3.4. Discussion

The methodology we used in Experiment 1 has resulted in the large proportions of participants accepting the priors given to them, not violating the independence of the causes before learning the effect, and correctly answering the final logic question suggesting that they did understand the causal structure and the parameters of the cover stories. Despite these encouraging improvements, our findings echo those of the extant literature as participants overall insufficiently explained away. This was reflected in both poor diagnostic reasoning and poor direct qualitative explaining away as well as in insufficient qualitative relational explaining away in all three groups. Quantitative relational explaining away was insufficient in $\text{Group}_{\text{LOW}}$ and $\text{Group}_{\text{MED}}$ and marginally sufficient in $\text{Group}_{\text{HIGH}}$. The sufficient quantitative relational explaining away in $\text{Group}_{\text{HIGH}}$ could be attributed to the small normative amount of explaining away in the high condition which makes it easier for participants in this conditions to sufficiently explain away compared to participants in the other two conditions.

Since the different priors lead to different amounts of the normative explaining away we have predicted that participants would explain away more in low priors condition than in both medium and high priors conditions, and that participants reasoning with medium priors conditions would explain away more than those reasoning with high priors. We have found that participants' quantitative responses only partially supported this prediction: only in diagnostic reasoning we have found that the difference $P(\mathrm{C_i} \mid \mathrm{E}) - P(\mathrm{C_i})$ is the highest in the low condition, followed by the medium and the high condition. This was not found in participants' responses to quantitative questions regarding both the direct and relational explaining away. Interestingly, however, we have found that the proportions of participants correctly answering the *qualitative* questions regarding diagnostic reasoning and both the direct and relational explaining away did directly correspond to the size of the *quantitative* difference between the two probabilities and the normative amount of explaining away (which is dictated by the priors), with the highest proportion of participants correctly answering these qualitative questions being in the low conditions, followed by the medium condition, with the smallest proportion of correct answers found in the high condition. This finding is lending support to a claim that people are sensitive to the size of the normative differences between the probabilities being compared: the greater the quantitative normative difference the greater the proportion of people who will correctly choose the normative qualitative direction of probability change between the two probability estimates. This, however, was not the case with the participants' quantitative estimates which could be attributed to our two hypotheses.

As predicted by the propensity interpretation hypothesis, we found that a significant proportion of participants reported that $P(\mathrm{C_i}) = P(\mathrm{C_i} \mid \mathrm{E}) = P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ in both qualitative and quantitative questions. Moreover, we found that this proportion was the highest when participants were reasoning with the cover story in which we expected the propensity interpretation to be the most pronounced (Coins cover story) and the lowest when participants reasoned with the cover story in which we expected the propensity interpretation to he the least pronounced (Dinner party), with the third cover story (Ball and containers) falling between. This is exactly what is predicted by the propensity hypothesis. Furthermore, the cross-tabulations showed that the propensity hypotheses accounted for over 50% of violations in quantitative diagnostic reasoning and relational explaining away and over 70% of violations in qualitative diagnostic reasoning and explaining away (both direct and relational).

Finally, regarding our diagnostic split hypothesis we found that a significant proportion of participants in the low and medium conditions did split the probability space between the two causes in diagnostic reasoning and assigned .5 probability to each cause with the hypothesis accounting for over 50% of violations in quantitative diagnostic reasoning and relational explaining away. However, as the proportion of participants was significantly lower in the high conditions, the diagnostic split hypothesis was only partly supported. These results may suggest that people split the probability space in diagnostic reasoning only when the update to the diagnostic split prediction from the priors is in the qualitatively normative direction, a notion that is further explored in Experiment 2. The cross-tabulations in Table 5 also pointed that correct quantitative diagnostic reasoning could be predictive for explaining away: participants who correctly answered the quantitative diagnostic reasoning questions also correctly answered questions related to both the direct explaining away and the quantitative logic question. This is an interesting finding on its own as it may suggest that the crucial part in explaining away is diagnostic reasoning and that understanding violation in diagnostic reasoning will possibly lead to understanding violations in explaining away.

Taken together, the two hypotheses accounted for about 78% of violations in quantitative diagnostic reasoning and quantitative relational explaining away. Given this and the other above-mentioned high percentages, we can conclude that the diagnostic split hypothesis and the propensity hypothesis were able to explain a significant amount of the observed insufficiency in explaining away.

4. Experiment 2

4.1. Motivations

In Experiment 1 the diagnostic split hypothesis had as a prediction .5 probability for each cause in diagnostic reasoning. However, it is not uncommon that people assign probability of .5 to events when they want to express their lack of confidence in their answer or when they want to express that they do not know what the answer is (see for example Fischhoff & Bruine De Bruin, 1999). So rather than following the diagnostic split strategy, an alternative explanation regarding Experiment 1 findings where some people gave .5 as their estimates to diagnostic reasoning questions, is that these people were expressing that they did not know the answers. The goal of Experiment 2 was to disentangle the two possibilities and further extend results of Experiment 1 to more than 2 causes. To do so, in Experiment 2 we prompted participants to reason with a 4-node common-effect CBN with three causes (see Figure 7).

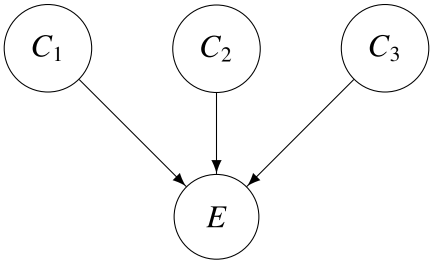

Figure 7: A common-effect CBN model with three causes.

In the CBN from Figure 7, assuming equal priors of all 3 causes and the deterministic set-up like in Experiment 1, the diagnostic split hypothesis would predict that $P(\mathrm{C}_1 \mid \mathrm{E}) = P(\mathrm{C}_2 \mid \mathrm{E}) = P(\mathrm{C}_3 \mid \mathrm{E}) = \frac{1}{3} \approx .33$. As .33 is sufficiently distinct from a .5 response that could also be a stand in for ‘I am not sure’ or ‘I do not know’, if people’s diagnostic reasoning judgments go to .33 that would suggest that these people do employ the diagnostic split strategy.

Another goal of Experiment 2 was to further test a prediction of the diagnostic split hypothesis whereby given high enough priors the split in the diagnostic reasoning would result in $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E})$ being lower than $P(\mathrm{C}_\mathrm{i})$ (as was the case in High condition in Experiment 1), which is opposite to the normative direction of the update where $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}) > P(\mathrm{C}_\mathrm{i})$. In Experiment 1 we found that only around 14% of participants’ estimates went down from .7 priors to .5 in diagnostic reasoning compared to half of participants’ estimates that went up from .2 priors to .5 in diagnostic reasoning. This suggests that people were significantly less inclined to reduce the probability of the causes in diagnostic reasoning. Experiment 2 was set to test this prediction in the context of three causes. If the results from Experiment 1 were replicated, then the diagnostic split hypothesis would need to be revised to account for small proportion of people who reduce the probability of causes in diagnostic reasoning.

4.2. Overview

Similarly to Experiment 1, we manipulated the priors of causes and presented participants with different cover stories. We again employed a deterministic set-up where the presence of at least one cause entailed the presence of the effect: $P(\mathrm{E} \mid \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3) = P(\mathrm{E} \mid \mathrm{C}_\mathrm{i}, \mathrm{C}_\mathrm{j}, {\sim}\mathrm{C}_\mathrm{k}) = P(\mathrm{E} \mid \mathrm{C}_\mathrm{i}, {\sim}\mathrm{C}_\mathrm{j}, {\sim}\mathrm{C}_\mathrm{k}) = 1$; and absence of all three causes entailed absence of the effect: $P(\mathrm{E} \mid {\sim}\mathrm{C}_1, {\sim}\mathrm{C}_2, {\sim}\mathrm{C}_3) = 0$. In this experiment, however, the priors were either either low, $P(\mathrm{C}_1) = P(\mathrm{C}_2) = P(\mathrm{C}_3) = .2$ or medium, $P(\mathrm{C}_1) = P(\mathrm{C}_2) = P(\mathrm{C}_3) = .5$. We deemed these two variations of priors to be sufficient to (i) disentangle the probabilistic split strategy predictions from an alternative mentioned above and (ii) further test the diagnostic split hypothesis on its prediction in the medium condition where $P(\mathrm{C_i} \mid \mathrm{E}) = .33 < .5 = P(\mathrm{C_i})$.

In this experiment we employed two cover stories from Experiment 1, one involving balls and containers, and one involving a dinner party. We did not use the cover story involving coin tossing since Experiment 1 findings suggested that participants reasoning within that cover story stayed significantly more at their priors when answering diagnostic reasoning questions compared to participants reasoning with the other two cover stories. As the primary goal of Experiment 2 is to distinguish between people giving .5 estimate to express their lack of confidence and the diagnostic split strategy, which required providing estimates different to the prior probabilities, to increase the power of Experiment 2 we did not include the cover story including coin tossing.

Further, since in Experiment 1 the tests regarding the propensity hypothesis were not significant between the balls and containers cover story and the dinner party cover story we have not directly tested the propensity hypothesis in Experiment 2. However, given that the propensity hypothesis has a clear prediction in Experiment 2, namely $P(\mathrm{C_i} \mid \mathrm{E}) = P(\mathrm{C_i})$ for $\mathrm{i} = \{1, 2, 3\}$, we again cross-tabulated the data to explore how much of the violation in diagnostic reasoning can be explained by the propensity hypothesis.

Given the new structure in Figure 7, in the balls and container cover story the three causes were now represented by three balls (binary variables assuming the value of either copper or rubber), randomly selected from three independent containers and placed on three gaps in an electric circuit. If at least one of the three balls was copper, a light bulb in the circuit (common effect) would turn on. In the dinner party cover story the three causes were represented by three individuals, Michael, Tom and Sam, and the common effect was represented by a fourth individual, Helen, who would drink wine only if at least one of the three aforementioned people brought wine to a party ('Helen' was a binary variable assuming the value of either ‘drinking wine’ or ‘not drinking wine’).

4.3. Methods

4.3.1. Participants and Design

A total of 119 participants (N_{MALE} = 39, 2 participants identified as ‘other’, M_{AGE} = 35 years). All participants were native English speakers who gave informed consent and were paid £1 for partaking in the present study, which took on average 8.25 minutes to complete.

A between-participant design was employed and participants were randomly allocated to one of 2 (cover story: ball containers, dinner party) × 2 (priors condition: low, medium) = 4 groups $N_{\text{BALL\_CONTAINERS\_LOW}} = 28$, $N_{\text{BALL\_CONTAINERS\_MED}} = 30$, $N_{\text{DINNER\_LOW}} = 32$, $N_{\text{DINNER\_MED}} = 29$).

4.3.2. Materials

Each of the groups was asked to complete an inference questionnaire ($N_{\text{QUESTIONS}}$ = 12), comprising of questions regarding priors and (unconditional) independence of causes, as well as reasoning questions relating to diagnostic reasoning and explaining away. For a full list of questions and the inferences these represented see Table 8. For diagnostic reasoning inferences, two questions were asked regarding the same inference, one in qualitative format (e.g. Q7) and one in quantitative format (e.g. Q8).

Each of the four groups ether reasoned with low or medium priors and was either presented the balls and containers cover story or the dinner party cover story from Experiment 1 now adapted to include the third cause. For full materials visit Open Science Framework, https://osf.io/aqjkp/.

4.3.3. Procedure

Like in Experiment 1, participants in each of the four groups were initially presented with the pertinent cover story and were given explicit information on the common-effect

Table 8: Inference types and questions found in the questionnaire for Experiment 2.

Question Number	Inference Type	Key Inferences	Question Type
1		$P(\mathrm{C}_1)$	Quantitative
2	**Priors**	$P(\mathrm{C}_2)$	Quantitative
3		$P(\mathrm{C}_3)$	Quantitative
4		$P(\mathrm{C}_2 \mid \mathrm{C}_1)$	Qualitative
5	**Independence**	$P({\sim}\mathrm{C}_3 \mid {\sim}\mathrm{C}_2)$	Qualitative
6		$P(\mathrm{C}_1 \mid {\sim}\mathrm{C}_3)$	Qualitative
7 , 8		$P(\mathrm{C}_1 \mid \mathrm{E})\text{-}R\text{-}P(\mathrm{C}_1)$	Qual. + Quant.
9 , 10	**Diagnostic Reasoning**	$P(\mathrm{C}_2 \mid \mathrm{E})\text{-}R\text{-}P(\mathrm{C}_2)$	Qual. + Quant.
11 , 12		$P(\mathrm{C}_3 \mid \mathrm{E})\text{-}R\text{-}P(\mathrm{C}_3)$	Qual. + Quant.

Note: -*R*- stands for ‘in relation to’.

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 in visual form (graphical representation). In order to ensure participants understood the structure, they 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 see Table 8). The questionnaire required participants to *sequentially* answer questions firstly regarding priors of causes, secondly independence of causes and finally regarding diagnostic reasoning about each cause. The graphical and textual details of the cover story were present on the same page as the relevant inference questions so participants could access these details at any point.

Questions marked as quantitative in Table 8 required participants to provide numerical estimates on a slider with a scale ranging from 0% to 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. $P(\mathrm{C}_2 \mid \mathrm{C}_1)$ given no knowledge of whether E is present or not. To investigate participants' diagnostic reasoning we employed both qualitative and quantitative question formats. For example, participants in groups reasoning with the balls and containers cover story, after finding out that the light bulb is on, were asked both a *qualitative* diagnostic reasoning question (e.g. Q7): "Does the probability that **Ball 1** is a copper ball **change** (compared to Q1, where you said: X%) after you find out that the light bulb turned on?" as well as a *quantitative* one: "What do you now think is the probability that **Ball 1** is a copper ball?". Additionally, diagnostic reasoning questions prompted participants to provide written explanations for their answers. All evidence (i.e. new states of cause or effect variables) was provided to participants both textually (e.g. in groups reasoning with balls container cover story: "You uncover the light bulb and find that it is turned on") as well as visually (as an updated graphical representation of the model). One again, the graphical and textual details of the cover story were present on the same page as the relevant inference questions so participants could access these details at any point.

4.4. Results

Participants' answers to all qualitative in the inference questionnaire are represented in Figure 8 and their responses to all quantitative questions are in Figure 9.

4.4.1. Overall Performance

As in Experiment 1, to test for a main effect of cover story and/or priors condition on participants' judgment accuracy throughout the inference questionnaire we initially

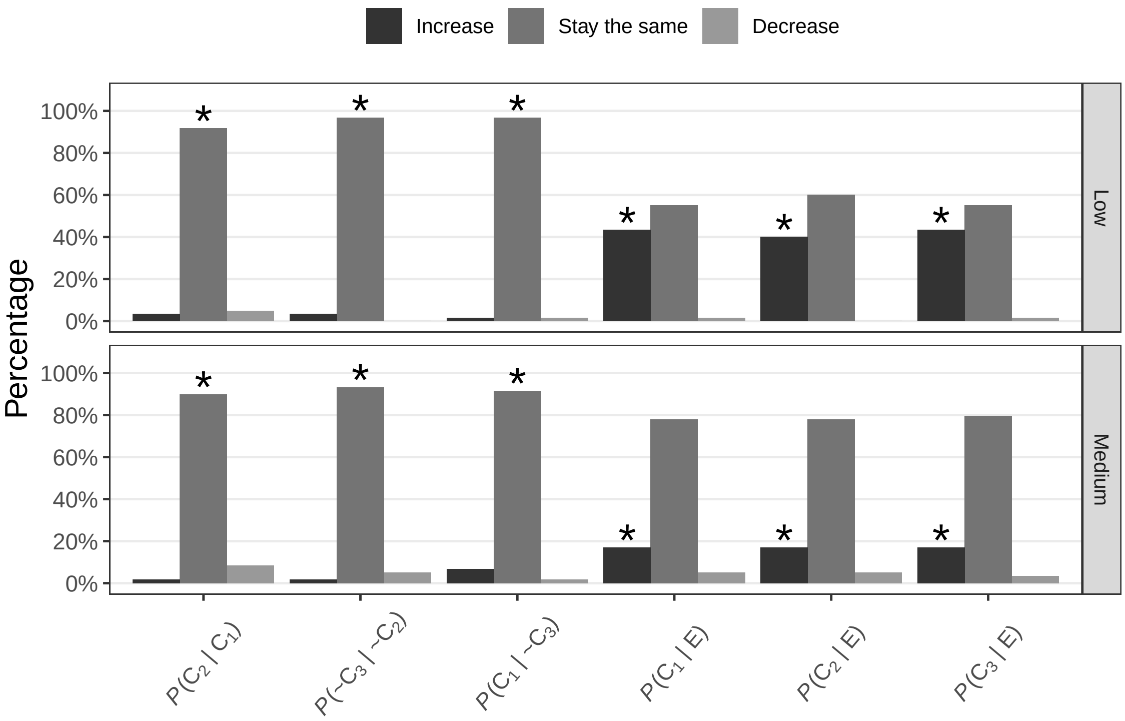

Figure 8: Distribution of participants' responses to qualitative questions in Experiment 2. Asterisks above the bars indicate normative answers.

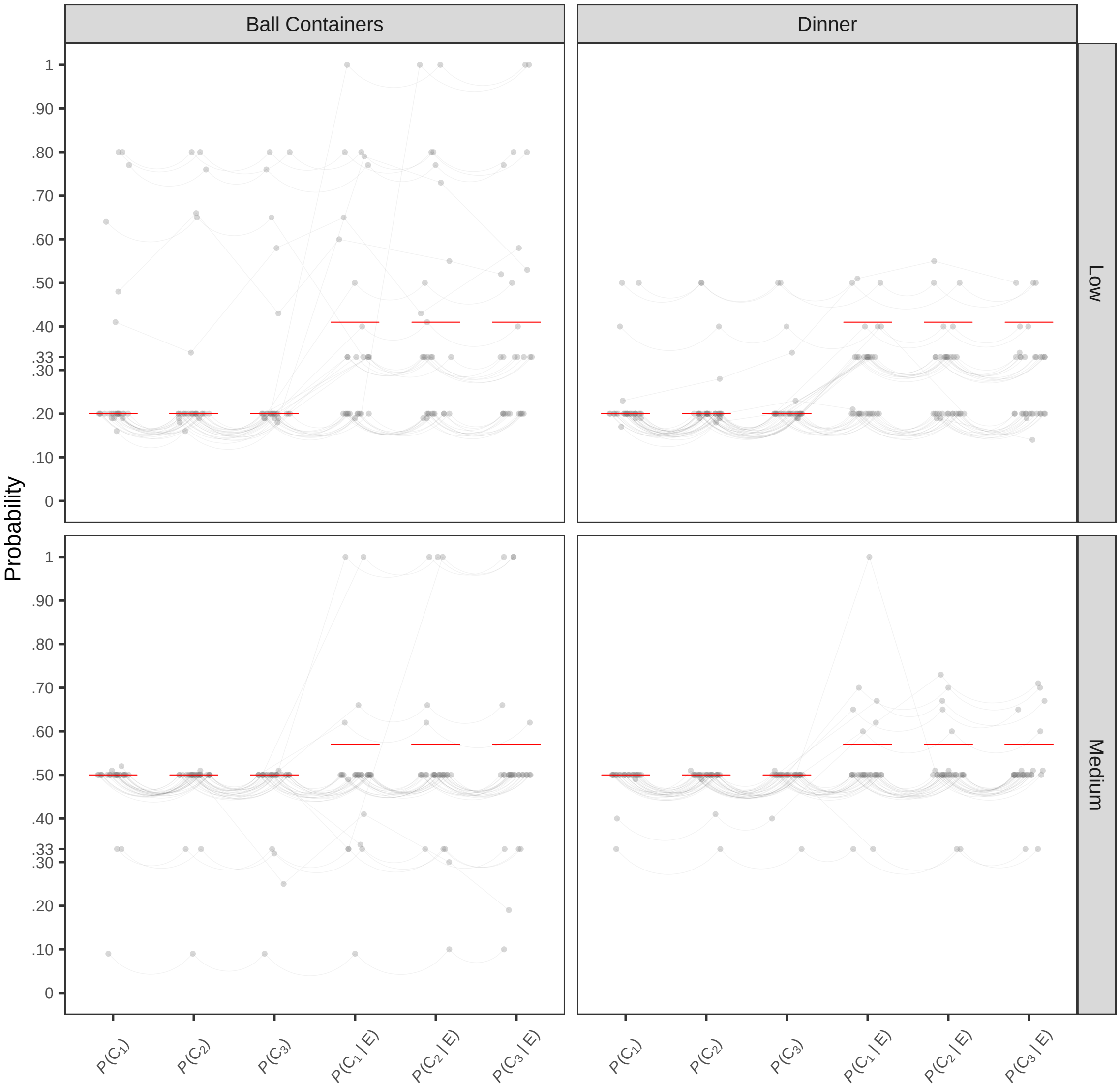

Figure 9: Participants' responses to quantitative questions in Experiment 2. Red horizontal lines are correct (normative) answers. Gray lines between data points depict how participants changed their probability estimates from one questions to another, with curved lines indicating that a participant did not change (within ± .02) their probability estimate.

re-coded all of participants' answers as being either correct (1) or incorrect (0). For all quantitative estimates, an answer was considered correct if it fell within ± .02 of the normative probability estimate. This allowed us to have a comparative measure of participants' accuracy in both qualitative and quantitative types of inferences. Subsequently, we combined each participants' coded response to the symmetrical pairs of inference into a single coded response: if a participant answered all three questions regarding priors correctly, the response was coded as 1; otherwise 0. Similarly was done for the questions regarding independence and qualitative and quantitative diagnostic reasoning. This left us with four coded question-types regarding: priors, independence, qualitative diagnostic reasoning, and quantitative diagnostic reasoning. For descriptive statistics of participant accuracy within each condition see Table 9 below.

Table 9: Descriptive statistics of participants' overall performance per group in Experiment 2.

Group	Proportion Correct Answers	95% CI
Balls Containers Low	.51	[.42, .59]
Balls Container Med	.47	[.39, .55]
Dinner Party Low	.54	[.47, .61]
Dinner Party Med	.49	[.44, .54]

To determine the effect of our manipulations on participants' overall performance throughout the task we built a GLM with binomial link function. The model had two fixed effects, Cover story and Priors, with a random intercept for each participant (there was no random slope for participant since Cover story and Priors vary between participants) and a random effect for question type. We found no main effect of Priors, $z = -0.89$, $p = .36$ and no main effect of Cover story, $z = -0.7$, $p = .49$. We also found no interaction between Cover story and Priors, $z = -0.03$, $p = .97$. Including the predictors (Cover story and Priors) in the model did not improve model fit ($\chi^2(3) = 1.32$, $p = .72$) compared to just having an intercept as a predictor. As our predictors were centered, this implied that the data grand mean fits the data no worse than the model which includes both predictors.

4.4.2. Accuracy

Given we found no effect of scenario or priors on participants' performance, we collapsed all conditions in order to obtain the following descriptives regarding participants' accuracy.

Prior probabilities. Collapsing across all conditions, 84% of participants correctly answered *all three* questions pertaining to the prior probabilities i.e. $P(\mathrm{C}_1)$, $P(\mathrm{C}_2)$ and $P(\mathrm{C}_3)$.

Independence. For a breakdown of the frequency of participants' answers to qualitative independence questions see Figure 8. Collapsing across conditions, 89% of participants correctly answered *all three* questions relating to independence (i.e. Qs 4, 5, and 6 in Table 8).

Diagnostic Reasoning. For a breakdown of the frequency of participants' answers to qualitative diagnostic reasoning questions see Figure 8. In regards to diagnostic reasoning, 26% of participants correctly answered all three *qualitative* diagnostic reasoning questions (Qs 7, 9, and 11 in Table 8) and only 2.5% of participants correctly answered all three *quantitative* diagnostic reasoning questions (Qs 8, 10, and 12 in Table 8).

4.4.3. Diagnostic split

In order to test the diagnostic split hypothesis we firstly collapsed the cover story condition and subsequently computed the proportion of participants who, having given the correct priors (± .02) for all three causes, updated the probabilities of $P(\mathrm{C}_1 \mid \mathrm{E})$, $P(\mathrm{C}_2 \mid \mathrm{E})$ and

$P(C_3 \mid E)$ to .33 (± .02) each. This proportion was 34% in group reasoning with low priors and 3.8% in group reasoning with medium priors. Chi-Square test of independence illustrated showed that these proportions significantly differed from each other, $\chi^2(1) = 13.48$, $p < .001$. Our findings replicate those of Experiment 1, as participants reasoning with low priors employed the diagnostic split strategy significantly more than participants who reasoned with medium priors.

Similarly to Experiment 1 analysis, we collapsed all data and cross-tabulated responses of participants who correctly answered all three priors questions. Table 10 illustrates that the diagnostic split hypothesis accounted for about 18% of violations in quantitative diagnostic reasoning.

4.4.4. Propensity interpretation

Although we have not explicitly tested the propensity hypothesis in this experiment, the cross-tabulations showed how much of the violations in diagnostic reasoning can be accounted for by this hypothesis. Table 10 shows that about 67% of the participants who failed *quantitative* diagnostic reasoning reasoned in line with the propensity interpretation (i.e. they provided estimates $P(C_i \mid E) = P(C_i)$ (± .02) for all three causes). Table 11 further shows that about 93% of the participants who failed *qualitative* diagnostic reasoning reasoned in line with the propensity interpretation (i.e. they responded with ‘stay the same’ for all three comparison between the priors and the diagnostic reasoning). These results suggest that the propensity hypothesis accounted for a significant proportions of failures in diagnostic reasoning.

4.4.5. Discussion

In Experiment 2 we found that the majority of participants accepted the priors given to them and did not violate the assumption of independence of causes prior to learning of the effect. These findings corroborate those of Experiment 1 and suggest that participants

Table 10: A cross-tabulation for correct/incorrect (yes/no) quantitative diagnostic reasoning as well as for in line/not in line with (yes/no) the diagnostic split hypothesis and the (quantitative) propensity hypothesis predictions in Experiment 2.

		Quantitative diagnostic reasoning	
		Yes	*No*
Diagnostic split	Propensity interpretation (quantitative)		
Yes	*Yes*	0	0
Yes	*No*	0	18
No	*Yes*	0	66
No	*No*	2	14

had a good understanding of the causal structure, parameters, and the cover story they were reasoning with. Despite this, we once again found that participants in all conditions performed poorly in diagnostic reasoning, especially when this was measured as accuracy of quantitative probability estimates.

In regards to our diagnostic split hypothesis, we found that it accounted for about 18% of violations in diagnostic reasoning. More specifically, we found that a significant portion of participants employed this strategy in the group reasoning with low priors, who increased their probabilities of $P(\mathrm{C_i})$ from .2 to .33. Disparately, this strategy was scarcely utilised by the groups reasoning with medium priors, who, according to the hypothesis would have had to decrease their prior probability estimates of each cause from .5 to .33. Our findings therefore strengthen the notion that the diagnostic split hypothesis is

Table 11: A cross-tabulation for correct/incorrect (yes/no) qualitative diagnostic reasoning as well as for in line/not in line with (yes/no) the (qualitative) propensity hypothesis predictions in Experiment 2.

	Qualitative diagnostic reasoning	
	Yes	*No*
Propensity interpretation (qualitative)		
Yes	0	67
No	28	5

dependent on the normative direction of the update from the priors. When the diagnostic split hypothesis predicts a value that is below that of the prior probability of the cause, then participants' behaviour does not follow the prediction. This is in accordance with the findings of Experiment 1 where we observed a dearth of participants who engaged in the diagnostic split strategy when reasoning with high priors ($P(\mathrm{C_i}) = .7$)). An intuitive explanation would be that as evidence is positively correlated with a cause, learning of the presence of the evidence (effect) would not *decrease* the probability of the cause. Overall findings from Experiment 2 solidify the presence of the diagnostic split hypothesis (in the normative direction of update) and serve to demonstrate that underlying participants' updating behaviour in Experiment 1 (attributing .5 to each cause) was not due to a lack of confidence or an unawareness of the task, but an engagement in a specific strategy.

Another updating behaviour that accounted for a large cluster of participants' data is encompassed by the propensity hypothesis. We found that about two thirds of the violations in quantitative and over 90% of violations in qualitative diagnostic reasoning can be explained by the propensity hypothesis. Although we have not explicitly tested the propensity hypothesis in Experiment 2 these proportions provide further empirical support for it.

Overall our findings show that the diagnostic split hypothesis and the propensity hypothesis are able to explain the vast majority of the violations in our data, thus suggesting that underlying pitfalls in diagnostic reasoning are pervasive, but could be accounted for by specific reasoning strategies.

5. General Discussion

Over the past few decades, causal Bayesian networks have been successfully utilised to build normative and descriptive accounts of various facets of human reasoning. Despite this, they have so far failed to account for people's behaviour when engaging in explaining away. Empirical work in psychological literature has repeatedly demonstrated that people violate the normative CBN model in numerous ways when carrying out explaining away inferences.

We carried out two experiments utilising a novel methodology to address the issues found in previous empirical studies of explaining away that arguably partly accounted for people's recurrent deviations from the normative model. For example, we explicitly stated the prior probabilities of the causes found in our model and re-elicited these from participants in order to ascertain that these were accepted. Moreover, we utilised relational qualitative and quantitative question formats to elicit probabilistic inferences from participants. This allowed us to assess people's accuracy in providing single point estimates as well as in detecting probabilistic changes in the model in a qualitative, more intuitive, fashion. This approach was 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 both experiments and across conditions, we found that a high proportion of participants answered correctly questions relating to priors, independence of causes as well as the final logic question.

The assumption of independence is often reported to be violated in the majority of studies that find insufficient explaining away (Rottman & Hastie, 2016; Mayrhofer & Waldmann, 2015; Rehder & Waldmann, 2017). Assuming the causes are independent before learning of the presence of the effect can be crucial since positive correlation between the causes can drastically reduce the normative amount of explaining away. Notably however, in both our experiments we found no violation of this assumption in any condition. All studies that reported a violation of the assumption of independence utilised quantitative questions to (unsuccessfully) elicit participants understanding of the independence of causes. Given the findings from our experiments and given encouraging finding from Rehder (2014a) who also employed a version of qualitative forced choice questions, we advocate that utilising qualitative questions to address this understanding might be a promising way forward.

In addition, in Experiment 1 we found that a large proportion of participants correctly answered the final logic question. This finding is important as it suggests that participants did understand the logical structure of the problem presented to them. However, some studies on explaining away reported a small percentage of participants as being able to solve questions pertaining to this inference. For instance, Rottman and Hastie (2016) report that less than 10% in Experiment 1a and only around 29% in Experiment 1b of responses correctly concluded that after learning the evidence, additionally learning that one causes did not occur means that the other one must have occurred (in their study they also had that $P(\mathrm{E} \mid {\sim}\mathrm{C}_\mathrm{i}, {\sim}\mathrm{C}_\mathrm{j}) = 0$ which implies that $P(\mathrm{C}_\mathrm{i} \mid \mathrm{E}, {\sim}\mathrm{C}_\mathrm{j}) = 1$).

Despite our encouraging findings regarding priors, independence, and logic, our main findings echoed those of the extant literature as participants in both experiments overall systematically violated the normative account of explaining away (Davis & Rehder, 2017;

Fernbach & Rehder, 2013; Morris & Larrick, 1995; Rehder, 2014a; Rehder & Waldmann, 2017; Rottman & Hastie, 2016; Sussman & Oppenheimer, 2011). In Experiment 1 pitfalls in relational explaining away comprised of both poor diagnostic reasoning and direct explaining away in both quantitative and qualitative questions. Further, participants' answers to quantitative inference questions were corresponding to different amounts of explaining away only in diagnostic reasoning. Notably however, our results suggested that the proportions of participants correctly answering the *qualitative* questions did directly correspond to the normative amount of explaining away, a fining that should further be explored. In addition, findings from both of our experiments allowed us to conclude that deviations from the normative model observed in our experiments could not be attributed to structural violations to the normative model (i.e. violations of the independence condition), as past literature intimated, but instead seem to arise, at least in part, from participants utilising certain sub-optimal reasoning strategies such as the diagnostic split and interpreting probabilities as propensities.

5.1. Diagnostic split

The findings of the two experiments suggest that some people do equally split the probability space between the two causes when engaging in diagnostic reasoning. As such, we found that a significant proportion of participants' answers aligned with predictions made by the diagnostic split hypothesis. Furthermore, Experiment 2 tested the strategy in the context of three causes and excluded an alternative explanation of the findings from Experiment 1 that posits that participants who provided .5 as an estimate in diagnostic reasoning were not driven by the diagnostic split strategy but rather trying to communicate that low confidence or an inability to respond to the question. However, the findings from Experiment 1 suggested that people were not willing to *decrease* the probability from the priors to the prediction of the diagnostic split hypotheses; they rather stayed at the priors in diagnostic reasoning. As this was further explored and confirmed in Experiment 2, we need to modify our diagnostic split hypothesis to account for this. The hypothesis then holds only when its predictions align with the qualitative predictions of the normative account: if, for example, the normative account implies that $P(\mathrm{C_i}) \leq P(\mathrm{C_i} \mid \mathrm{E})$ for $1 \leq i \leq n$, then the diagnostic split hypothesis predicts that $P(\mathrm{C_i} \mid \mathrm{E}) = \frac{1}{n}$ when the priors are equal, the set-up is deterministic, and $P(\mathrm{C_i}) \leq \frac{1}{n}$.

Crucially, through the use of cross-tabulations we were able to illustrate that in Experiment 1 adopting a diagnostic split strategy accounted for 51% of observed deviations in both quantitative diagnostic reasoning and quantitative relational explaining away. In Experiment 2 approximately 18% of violations in quantitative diagnostic reasoning could be attributed to a diagnostic split strategy. Ultimately this allowed us to support the notion that this strategy contributes significantly to the observed violations of explaining away.

So far we have only explored the diagnostic split hypothesis in a deterministic set-up where the presence of at least one cause entails the presence of an effect and where the effect cannot occur when none of the causes are present; or where after learning the effect one of the causes (or both) must have happened, i.e. the causes are exhaustive. However, there is evidence that the hypothesis also applies to less deterministic contexts. For instance, Rottman and Hastie (2016) found spikes in data around the .5 probability from their Experiment 1 where the priors were the same for the two causes and the causes became exhaustive after learning the effect, but a presence of a cause did not entail the presence of the effect. Whether the diagnostic split hypothesis holds in the context where a presence of a cause does not entail the effect (but the causes are still exhaustive after learning the effect) should be explored in future work.

5.2. Propensity interpretation

The findings from both experiments also suggest that a large number of participants remained at the priors when answering diagnostic reasoning and direct explaining away questions. Moreover, Experiment 1 showed that the proportions of participants who remain at the priors are different in the three cover stories with the proportion of participants being the largest in the cover story where we argued the propensity interpretation is the most pronounced, the smallest in the cover story with the least pronounced propensity interpretation, and in between in the third cover story. These findings fit the predictions of the propensity interpretation, thus providing support for it. Further, we have found that the propensity hypothesis is able to account for a significant amount of insufficiency in explaining away. The cross-tabulations in Experiment 1 showed that the propensity interpretation was able to account for 53% of violations in both the quantitative diagnostic reasoning and quantitative relational explaining away and over 70% of violations in qualitative diagnostic reasoning, direct and relational explaining away. In Experiment 1 we found that the propensity hypothesis could account for over 90% of failures in qualitative diagnostic reasoning. These percentages allowed us to find support for our theory that adopting this interpretation of probability can significantly account for violations of patterns of inferences within explaining away.

The prediction of the propensity interpretation, however, are not limited to situations exhibiting explaining away. It also applies to any contexts where probabilities could be interpreted as established propensities, especially if they include causal-probabilistic elements. These include common-effect structures in general (not just those exhibiting explaining away), but also common-causes and chain structures as well as simple two-node cause-effect structures. Specifically, in simple two-node structures the propensity interpretation could explain adherence to the prior and conservatism in belief updating, which seem to be often found in studies employing paradigms where probabilities are well-defined stochastic properties of an environment (Erev, Wallsten, & Budescu, 1994). This is particularly interesting as the propensity interpretation's prediction in the two-node cases are in direct opposition to the well-known base rate neglect where people partially or completely ignore the priors of causes (Barbey & Sloman, 2007; Eddy, 1982; Gigerenzer & Hoffrage, 1995; Tversky & Kahneman, 1974). The situations where we think that the propensity interpretation (or anchoring at the base rate) will be more pronounced than the base rate neglect are those that are characterized by (i) a deterministic set-up, (ii) clearly defined stochastic properties of (physical) systems, and (iii) clear causal-mechanistic relations between the parts of the (physical) system or between multiple physical systems. The situations involving social interactions where relations are less deterministic or less clearly related in a causal-mechanistic way would, in our opinion, be more prone to people neglecting the priors. These should be explored in the future work.

Finally, we would like to touch on the normative status of the propensity hypothesis. As the propensity interpretation is one of the interpretations of probability one might think that it should agree with the normative account. However, as mentioned in the introduction, the problems for propensity interpretation have been raised, such the Humphreys's paradox, that challenge the idea that it can be reconciled with the axioms of probability which are the bedrocks of the normative account. Furthermore, the propensity interpretation's predictions that the probability of a cause does not change in light of an effect (and in light of additionally learning the other cause has obtained) goes against the Bayesian updating (also known as 'conditionalization') which as a consequence has that an agent following the propensity interpretation in this case is not uniquely minimising the inaccuracy of its beliefs when that inaccuracy is measured with a proper scoring rule such as the Brier score (for details see Pettigrew, 2016). On the other hand, in Section 2.2 we have seen that, under certain conditions, the propensity interpretation can be a good approximation of the normative account. It is, however, outside the scope of this paper to argue for or against the normative status of the propensity interpretation. We simply find that the propensity interpretation is a good descriptive account of the findings on explaining away.

5.3. Limitations

A few important limitations of the current study are in order. First, in both experiments priors and conditional probabilities have been communicated textually and graphically to participants. We have not explored whether our findings replicate when participants are presented with learning data. Since with learning trials priors would not be 'established' but inferred from data and function as estimates of priors, we expect the propensity interpretation to be less pronounced. As a consequence we would expect less participants to stay at the priors in diagnostic reasoning and explaining away compared to the findings in the current study. However, we would still expect participants to split the probability space in diagnostic reasoning as per the diagnostic split. This is supported by Rottman and Hastie (2016) who utilized learning trials in their study.

Second, we have only considered explaining away in a deterministic set-up. Admittedly this is fairly limiting from a perspective of the ecological validity of our findings. We proposed further avenues of research with respect to this limitation and have also argued that we expect to find similar results with respect to both hypotheses even in less deterministic set-ups.

Third, in both experiments we have used the same quantitative response scale that promoted participants enter a number between 0 and 100 eliciting from them the probability with which the participants believed a certain event (a coin landing Heads) would happen. However, other response scale formats are available. For instance, a frequency format response scale (Gigerenzer & Hoffrage, 1995) would ask participants to provide the number of coins (that are like the coins in the cover story) that they would expect to land Heads given that the light bulb turned on out of the total number of these coins that land Heads.

The primary reasons we have not used, for instance, the frequency format response scale is that (i) given the events in our cover stories are token events that had occurred only once (Coin 1 landed once, Ball 1 was picked for a container once, and Michael is coming to a party at a particular location on a particular time) the frequency format (which refers to a frequency with which an outcome occurs in a sequence of similar events) would not have fit well with the single occurrences of token events and (ii) eliciting frequencies from participants would, we believe, steer them away from the propensity probability interpretation towards the frequency interpretation (which is out of the scope of the current paper) thus reducing the power of our experiments. However, further studies should explore different response scales formats, such as the frequency format, that would arguably put more emphasis on different probability interpretation, like the frequency probability interpretation. This would allow for a further exploration of the role of probability interpretations in explaining away and causal reasoning in general.

Fourth, we recognize that in some cases it may not be straightforward to determine whether probabilities are interpreted as propensities or in some other way. There is no normative computational procedure that could tell how probabilities should be interpreted. One can only provide arguments for or against a certain interpretation and rely on these when testing in contexts embodying a certain interpretation. Most difficulties arise when discussing possible borderline cases. For instance, some philosophers have argued that probabilities in medical contexts, which are often employed in psychological experiments, are on the border between epistemological and objective interpretations and could lean either way (Gillies, 2000a). This, however, does not render empirical exploration of people's intuitions about different probability interpretations futile. As long as there is a sufficient consensus regarding how clear-cut are the specific contexts for testing particular interpretations, one should be on a safe side employing these in their empirical studies. Even in cases that are not clear-cut one can employ different elicitation methods to test different interpretations, e.g. one could use different phrasings of questions (c.f. Ülkümen et al., 2016).

5.4. Conclusion

In our experiments we have replicated findings in the extant literature reporting insufficient explaining away. We have also shown that this insufficiency is not due to violations of the independence assumption, as is sometimes suggested. Instead, we found that the insufficiency can largely be accounted by the two hypotheses, i.e. the diagnostic split strategy and propensity probability interpretation. Although we explored explaining away only in a deterministic context, we regard this context as a good starting point from which further research avenues emerge where the robustness of the two hypotheses could be addressed.

Appendix A

Here we show that Inequality 1 holds even when one or both causes in the explaining away situation are inhibitory. First, notice that $P(\mathrm{E} \mid \mathrm{C_i}) < P(\mathrm{E})$ if and only if $P(\mathrm{C_i} \mid \mathrm{E}) < P(\mathrm{C_i})$ and $P(\mathrm{E} \mid \mathrm{C_i}) > P(\mathrm{E})$ if and only if $P(\mathrm{C_i} \mid \mathrm{E}) > P(\mathrm{C_i})$ (proofs omitted). Then we have that:

$$\begin{aligned}
P(\mathrm{C_i} \mid \mathrm{E}) &= \frac{P(\mathrm{C_i})\ \sum_{C_j} P(C_j)\, P(\mathrm{E} \mid \mathrm{C_i}, C_j)}{\sum_{C_i,C_j} P(C_i)\, P(C_j)\, P(\mathrm{E} \mid C_i, C_j)} \\
P(\mathrm{C_i} \mid \mathrm{E}) - P(\mathrm{C_i}) &= P(\mathrm{C_i}) \frac{\sum_{C_j} P(C_j)\, P(\mathrm{E} \mid \mathrm{C_i}, C_j) - \sum_{C_i,C_j} P(C_i)\, P(C_j)\, P(\mathrm{E} \mid C_i, C_j)}{\sum_{C_i,C_j} P(C_i)\, P(C_j)\, P(\mathrm{E} \mid C_i, C_j)} \\
&= P(\mathrm{C_i})\, P(\sim\mathrm{C_i}) \frac{\sum_{C_j} P(C_j)\, P(\mathrm{E} \mid \mathrm{C_i}, C_j) - \sum_{C_j} P(C_j)\, P(\mathrm{E} \mid \sim\mathrm{C_i}, C_j)}{\sum_{C_i,C_j} P(C_i)\, P(C_j)\, P(\mathrm{E} \mid C_i, C_j)} \\
&= P(\mathrm{C_i})\, P(\sim\mathrm{C_i}) \frac{A - B}{\sum_{C_i,C_j} P(C_i)\, P(C_j)\, P(\mathrm{E} \mid C_i, C_j)}, \text{ where} \\
A &:= P(\mathrm{C_j}) \left[P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j}) - P(\mathrm{E} \mid \sim\mathrm{C_i}, \mathrm{C_j})\right], \text{ and}
\end{aligned}$$

$$B := P(\sim\mathrm{C_j}) \left[P(\mathrm{E} \mid \sim\mathrm{C_i}, \sim\mathrm{C_j}) - P(\mathrm{E} \mid \mathrm{C_i}, \sim\mathrm{C_j}) \right].$$

Therefore, $P(\mathrm{E} \mid \mathrm{C_i}) < P(\mathrm{E})$ if and only if $A < B$ and $P(\mathrm{E} \mid \mathrm{C_i}) > P(\mathrm{E})$ if and only if $A > B$. To see how this result corresponds to explaining away, we write again Inequality 1:

$$P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j})\, P(\mathrm{E} \mid \sim\mathrm{C_i}, \sim\mathrm{C_j}) < P(\mathrm{E} \mid \mathrm{C_i}, \sim\mathrm{C_j})\, P(\mathrm{E} \mid \sim\mathrm{C_i}, \mathrm{C_j})$$

It is easy to see that when, for instance, $P(\mathrm{E} \mid \mathrm{C_1}, \mathrm{C_2}) = 0$, $P(\mathrm{E} \mid \mathrm{C_i}, \sim\mathrm{C_j}) = 1$ and $P(\mathrm{E} \mid \sim\mathrm{C_1}, \sim\mathrm{C_2}) = 1$, both causes are inhibitory as $P(\mathrm{C_i} \mid \mathrm{E}) < P(\mathrm{C_i})$ for both causes, but Inequality 1 is still satisfied. Similarly, assuming the priors are equal, when $P(\mathrm{E} \mid \mathrm{C_1}, \mathrm{C_2}) = P(\mathrm{E} \mid \sim\mathrm{C_1}, \sim\mathrm{C_2}) = 0$, $P(\mathrm{E} \mid \mathrm{C_1}, \sim\mathrm{C_2}) = 1$ and $P(\mathrm{E} \mid \sim\mathrm{C_1}, \mathrm{C_2}) = .1$, then cause $\mathrm{C_1}$ is generative ($P(\mathrm{C_1} \mid \mathrm{E}) > P(\mathrm{C_i})$) but cause $\mathrm{C_2}$ is inhibitory ($P(\mathrm{C_2} \mid \mathrm{E}) < P(\mathrm{C_2})$). Nonetheless, Inequality 1 remains satisfied.

Appendix B

Here we show that including participants' average estimates regarding the independence of C_1 and C_2 from Rottman and Hastie (2016, Experiment 1b) in the normative model leads to the explaining away effect not being normatively warranted.

To perform the calculations we assume that $P(\mathrm{C_i}) = .25$, $P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j}) = .75$, $P(\mathrm{E} \mid \mathrm{C_i}, \sim\mathrm{C_j}) = P(\mathrm{E} \mid \sim\mathrm{C_i}, \mathrm{C_j}) = .5$, $P(\mathrm{E} \mid \sim\mathrm{C_i}, \sim\mathrm{C_j}) = 0$, as is stated in the study. There is some empirical support that participants accepted $P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j}) = .75$ (although there is a lot of variation in participants' estimates). There is, however, no data reported on whether participants accepted other parameters. Lastly, form the study we have that participants average estimates regarding independence are $P(\mathrm{C_i} \mid \mathrm{C_j}) = .45$ and $P(\mathrm{C_i} \mid \sim\mathrm{C_j}) = .35$.

$$P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j}) = \frac{P(\mathrm{C_i}, \mathrm{C_j}, \mathrm{E})}{P(\mathrm{C_j}, \mathrm{E})} = \frac{P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j})\, P(\mathrm{C_i} \mid \mathrm{C_j})\, P(\mathrm{C_j})}{P(\mathrm{E} \mid \mathrm{C_j})\, P(\mathrm{C_j})} = \frac{P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j})\, P(\mathrm{C_i} \mid \mathrm{C_j})}{P(\mathrm{E} \mid \mathrm{C_j})}$$

$$= \frac{P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j})\, P(\mathrm{C_i} \mid \mathrm{C_j})}{\sum_{C_i} P(\mathrm{E} \mid C_i, \mathrm{C_j})\, P(C_i \mid \mathrm{C_j})} = \frac{.75 \times .45}{.75 \times .45 + .5 \times .55} \approx .55$$

$$\begin{aligned} P(\mathrm{C_i} \mid \mathrm{E}) &= \frac{P(\mathrm{C_i}, \mathrm{E})}{P(\mathrm{E})} = \frac{P(\mathrm{E} \mid \mathrm{C_i}) P(\mathrm{C_i})}{P(\mathrm{E} \mid \mathrm{C_i}) P(\mathrm{C_i}) + P(\mathrm{E} \mid \sim\mathrm{C_i}) P(\sim\mathrm{C_i})} \\ &= \frac{P(\mathrm{C_i}) \sum_{C_j} P(\mathrm{E} \mid \mathrm{C_i}, C_j)\, P(C_j \mid \mathrm{C_i})}{P(\mathrm{C_i}) \sum_{C_j} P(\mathrm{E} \mid \mathrm{C_i}, C_j)\, P(C_j \mid \mathrm{C_i}) + P(\sim\mathrm{C_i}) \sum_{C_j} P(\mathrm{E} \mid \sim\mathrm{C_i}, C_j)\, P(C_j \mid \sim\mathrm{C_i})} \\ &= \frac{.25 \times (.75 \times .45 + .5 \times .55)}{.25 \times (.75 \times .45 + .5 \times .55) + .75 \times (.5 \times .35 + 0)} \approx .54 \end{aligned}$$

Therefore, as $P(\mathrm{C_i} \mid \mathrm{E})$ and $P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ are very close to each other, the amount of explaining away is negligible with slightly going in the opposite direction to explaining away.

References

Anderst, J. D., Carpenter, S. L., & Abshire, T. C. (2013). Evaluation for bleeding disorders in suspected child abuse. *Pediatrics*, *131*(4), e1314–e1322.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological review*, *104*(2), 367–405.

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

Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman & P. Slovic (Eds.), *Judgment under uncertainty: Heuristics and biases* (Vol. 8, pp. 249–267). Cambridge University Press.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological review*, *101*(3), 519–527.

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

Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty–fifty = 50%? *Journal of Behavioral Decision Making*, *12*(2), 149–163.

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.

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

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, *102*(4), 684–704.

Gillies, D. (2000a). *Philosophical theories of probability*. London: Routledge.

Gillies, D. (2000b). Varieties of propensity. *The British journal for the philosophy of science*, *51*(4), 807–835.

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*.

Humphreys, P. (1985). Why propensities cannot be probabilities. *The philosophical* *review*, *94*(4), 557–570.

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.

Khemlani, S. S., & Oppenheimer, D. M. (2011). When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological bulletin*, *137*(2), 195.

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.*

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

Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 433–458). Oxford: Oxford University Press.

Meeker, W. Q., Hahn, G. J., & Escobar, L. A. (2017). *Statistical intervals: A guide for practitioners and researchers* (Vol. 541). Hoboken, NJ: John Wiley & Sons.

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 Upper Saddle 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.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.

Pilditch, T. D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological Science*. Retrieved from https://doi.org/10.1177/0956797618818484

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

Rehder, B. (2011). Reasoning with conjunctive causes. In *Proceedings of the cognitive science society* (Vol. 33).

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* (Vol. 33).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. Retrieved from http://www.jstor.org/stable/1738360

Tversky, A., & Kahneman, D. (1977). *Causal schemata in judgments under uncertainty*. Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a056667.pdf

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.

View publication stats