Explanation and argument

Explanation and Argument

Marko Tešić

2020

PhD Thesis

Department of Psychological Sciences

Birkbeck, University of London

Declaration

This thesis is the result of my own work. All collaborative aspects are explicitly acknowledged in the text.

Abstract

In this thesis I address the relationship between arguments and explanations. In particular, I consider three notions of explanation and the way they relate to arguments. I argue that arguments and explanations should be considered in tandem rather than in isolation. I provide support for this contention throughout the thesis. Specifically, I show how research on argumentation includes considerations of explanation and informs useful distinctions regarding the different notions of explanation (Chapter 1). I then explore the close relationship between arguments and explanations on a specific pattern of reasoning called 'explaining away' where I show how explanations can affect the strength of an argument (Chapter 2). In Chapter 3 I provide further theoretical background regarding the different notions of explanations and I discuss factors that constitute 'good' explanations. These factors are then explored with respect to the notion of explanations as inference processes and in the context of arguments viewed as causal Bayesian networks. In Chapter 4 I focus on one of the aspects of explanations when embedded in a social context. Here, I take a concern that has been prominent in recent argumentation work, namely the role of the argument source, and pursue it in the context of explanation by examining how the provision of explanations affects the perceived reliability of the explainer. Finally, in Chapter 5 I summarize the results of the thesis and discuss the implications and potential directions for future research.

Acknowledgements

I would like to thank my supervisor Ulrike Hahn for her enduring and persistent support and encouragement. I would also like to thank my second supervisor David Lagnado for his guidance and support. Many thanks to Alice Liefgreen who was a collaborator on three experiments in Chapter 2. I am grateful to the BARD project, the Alexander von Humboldt Foundation, and the Department of Psychological Sciences, Birkbeck for their financial support. Finally, I thank my friends and family for their support, especially during the tough times.

Contents

List of Figures

List of Tables

1

Introduction

People argue. They put forward arguments in virtually all spheres of life. Ranging from everyday situations (the bacon is burnt because it was cooked for too long), over political contexts (think of Brexit), to scientific ones (arguing for the existence of gravitational waves), and a plethora of other domains, people devise arguments to convince others (and/or themselves) (Hahn & Oaksford, 2012; Mercier & Sperber, 2011). It is then important to consider what constitutes a 'good' argument. Why are some arguments better than others? What

are the standard(s) for evaluating the quality of arguments? A lot of effort has been put into addressing these questions, with answers coming from different fields such as philosophy, psychology, linguistics, cognitive science, computer science (Hahn, 2020).

Explanations, like arguments, are also ubiquitous. They are sought by children and adults, in everyday and professional contexts, and are an integral part of science. We ask ourselves and others why certain events have happened: why is the road closed?; why is the bacon burnt?; why is the dog barking?; why is the Zoom meeting not starting?; why is Amazon recommending this book to me? A doctor may ask why a child is in pain. A lawyer may ask why a suspect is guilty of the crime that they've been accused of. A physicist may ask why general relativity is a good explanation of the changes in orbits of a binary pulsar. The propositions that address these requests are explanations (Lombrozo, 2012). For instance, an explanation for why a child is in pain could be that it has pulled a muscle; and an explanation for why a Zoom meeting has not started could be that the host has forgotten that there is one. The important question regarding explanations is what constitutes a 'good' explanation. Why are some explanations seem better than others? Are there any features that are markers of these good explanations? Answers to these questions have also come from a variety of fields, including philosophy, psychology, cognitive science, and computer science (Lombrozo, 2012; Tešić & Hahn, in press).

Ubiquity and prevalence are not the only features that are shared between arguments and explanations. Indeed, arguments and explanations are sometimes thought of as two sides of the same coin (Hempel, 1965). They both may have the same form and both provide reasons to answer 'why' question (Hahn,

2011). They are both exemplified in reasoning schemes such as the inference to the best explanation (Harman, 1965). Despite their similarities, arguments and explanations have often been studied in isolation.

In this thesis I study arguments and explanations in tandem. I explore (i) some of the impacts of explanations on the confidence we have in certain claims, distinctly a feature of arguments, (ii) how we explain arguments, and (iii) how some of the aspects of arguments such as their social nature translate into explanations and influence the effects of explanations. The contention of this thesis is that studying arguments and explanations in tandem rather than in isolation enriches our understanding of both, not least because the aspects of one can translate into the other and suggest fruitful research avenues.

In this introductory chapter I introduce arguments and explanations and their relationship. Specifically, I look into what makes a good argument and how some of the features of arguments translate in the context of explanations.

1.1 Explicating arguments

The main goal of argumentation is changing someone's beliefs about a particular standpoint (Hahn, 2011; Van Eemeren, Grootendorst, Johnson, Plantin, & Willard, 2013). However, even given this unique goal there are different ways to understand arguments. Much of the philosophical and psychological research on argumentation explicates arguments in terms of rational debate or an activity of reason where propositions are put forward to justify a specific standpoint from a perspective of rational evaluation criteria (Hahn & Oaksford, 2012; Van Eemeren et al., 2013). The emphasis here is on rationality and the rational evaluation criteria where belief changes are due to arguments being

in line with the rational criteria, implying that the belief change was rational as well.

Another way to understand arguments is through their actual ability to change beliefs and persuade people, regardless of whether the argument changing the beliefs are deemed rational by a particular rationality criterion or not. This way of explicating arguments is relevant to the psychological research on persuasion (e.g. Maio, Haddock, & Verplanken, 2018).

In this thesis, however, I adopt the former notion of arguments where the stress is on the rationality and the rational status of arguments. The idea that arguments have a rational status implies that there is (are) a standard(s) of rationality against which one can compare the arguments to evaluate their rational status. These standards of rationality, or normative frameworks as they are also called, allow us to judge whether the change in belief that resulted from there being an argument for a particular standpoint was rational or not. I next describe two of these normative frameworks.

1.1.1 Classical logic as the standard of argumentation

Consider the following the following argument:

All pigs have a snout.
Snowball is a pig.
Therefore:
Snowball has a snout.

From the perspective of classical logic this argument is valid: namely, it is impossible for the conclusion ('Snowball has a snout.') to be false and the

premises (a collection of sentences preceding the conclusion, i.e. 'All pigs have a snout.' and 'Snowball is a pig.') to be true. If we take logical validity as the standard of rationality against which we evaluate the quality of arguments, then the above argument would qualify as a good argument.

Logical validity, however, is a more general concept that deems certain arguments valid not because of the content of the argument, but because of the structural form the argument takes. The above argument is valid as it is a particular instantiation of a valid argument form:

All x have Y.

z is x.

Therefore:

z has Y.

Any argument that has the above form would be a valid argument, regardless of the content of that argument.

The lack of logical validity is, thus, considered to be a marker of bad arguments (Hahn, 2020). Some of the famous logical 'fallacies' (Hamblin, 1970) are often labels as 'fallacies' since they lack logical validity.

However, intuitively it is neither the case that all good arguments are valid arguments nor that all valid arguments are good arguments. For example, if after a number of laboratory tests one says

> The drug is safe because we have no evidence that it is not.

we would consider this a good argument (Hahn & Oaksford, 2007). However, the argument is not valid: it is possible that the conclusion ('the drug is safe')

is false and the premise ('no toxic effects of the drug were found in any of the tests') is true. Similarly, there are valid arguments that we would not consider as good arguments. For instance, we would not consider the following argument as a good argument:

> To allow everyone a freedom of speech is advantageous to the State, because it is highly conducive to the interest of the State that each individual enjoys a freedom of speech. (adapted from Hansen, 2015)

The argument, however, is valid—the premise and the conclusion have the same meaning, so whenever the premise is true, the conclusion is also true (for a discussion on circular arguments see Hahn, Oaksford, & Corner, 2005; Hahn & Oaksford, 2007). Logical validity, thus, does not seem to capture the intuitions we have about good arguments. The above examples further suggest that the quality of arguments does not depend just on their formal structure, but also on the content of arguments.

The limitations of classical logic as a normative standard for argument evaluation was already extensively discussed by Toulmin (1958/2003). To overcome these limitations, he proposed an influential framework that proved useful in identifying the different components of an argument (see Chapter 3 for further details). Nonetheless, Toulmin's framework provides little guidance on what the evaluating criteria for arguments are and in what way the content of arguments makes a difference in terms of argument quality, something that seems lacking in the logic framework as well (Hahn, 2020).

1.1.2 Bayesian argumentation

We have seen that focusing only on the structural relationships of arguments does not seem to capture their quality. Thus, a framework that goes beyond structural relationships whilst accounting for the content of arguments is needed. The Bayesian probabilistic framework of argumentation has been specifically developed with a goal to account for the effects of the *content* of arguments on their quality (Hahn & Oaksford, 2006, 2007, 2012; Hahn & Hornikx, 2016; Hahn, 2020). The general idea is that argument evaluation is akin to evidence learning: different arguments, just like different pieces of evidence, can have differential impact on our beliefs depending on how strong an argument or a piece of evidence is. The greater the impact of an argument on our beliefs the stronger the argument.

The two cornerstones of Bayesian argumentation (and Bayesianism in general, see Hájek & Hartmann, 2010; Hartmann & Sprenger, 2011) are the following. First, probabilities represent degrees of belief. The probability that some proposition h is true (i.e. $P(h)$) is representing our degree of belief in h being true. One of the early consequences of this frameworks is that people can assign different probabilities to the same propositions. If two people have different degrees of belief in a certain proposition, then they will assign different probabilities to that proposition. Second, people should change their beliefs about h (i.e. $P(h)$) in light of evidence e by conditionalizing on e. That is, one's new degree of belief in h (i.e. $P_{new}(h)$, also known as 'the posterior probability of h') is one's old degree of belief in h given evidence e (i.e. $P_{old}(h \mid e)$). Bayes's

theorem then tells us how to compute $P_{old}(h \mid e)$:

$$P_{old}(h \mid e) = \frac{P_{old}(h)\, P_{old}(e \mid h)}{P_{old}(h)\, P_{old}(e \mid h) + P_{old}(\sim h)\, P_{old}(e \mid \sim h)} \tag{1.1}$$

In the context of argumentation, e could represent a reason for believing the proposition/claim/decision h; or in other words, e could be viewed as a premise (or a set of premises) and h as a conclusion of an argument. How good an argument from e to h is depends then on the extent to which e increases one's degree of belief in h: whenever $P_{new}(h)(= P_{old}(h \mid e)) > P_{old}(h)$, then the premise(s) e provide some support for the conclusion h. The *strength* of an argument could then be defined as the posterior probability of the conclusion $P_{new}(h)$. To compare the strength of arguments across different claims one could specify the *force* of an argument, which could be measured in terms of the ratio of $P_{old}(e \mid h)$ and $P_{old}(e \mid \sim h)$ (also called 'the likelihood ratio') (Hahn & Oaksford, 2007), or using some other measures of confirmation (Fitelson, 1999; Tentori, Crupi, Bonini, & Osherson, 2007).

The framework of Bayesian argumentation has been successfully applied in different domains of argumentation. It has been particularly useful in explicating the 'fallacies' of argumentation, pointing to the aspects of these 'fallacies' that affect their argument quality, which often varies in different contexts (Bhatia & Oaksford, 2015; Corner, Hahn, & Oaksford, 2011; Haigh, Wood, & Stewart, 2016; Hahn & Oaksford, 2006, 2007; Hahn, 2011; Hahn & Hornikx, 2016; Harris, Hsu, & Madsen, 2012; Harris, Corner, & Hahn, 2013; Hoeken, Timmers, & Schellens, 2012; Hsu, Horng, Griffiths, & Chater, 2017; Jarvstad & Hahn, 2011).

1.1.2.1 The argument from sign

Another approach for evaluating the quality arguments with a specific look at the content of arguments is the scheme-based approach (for an overview of this approach see Walton, Reed, & Macagno, 2008). This approach has both a descriptive and a normative component. The descriptive component aims to identify different types of arguments or argument schemes. The normative component seeks to formulate appropriate 'critical questions' for each argument scheme. These critical questions then provide scheme-specific norms in the sense that the quality of a specific argument is dependent on the responses an arguer gives to these questions.

I describe both the scheme-based and the Bayesian argumentation approach in explicating an argument scheme called 'the argument from sign'. The argument from sign has the following form (Walton et al., 2008):

A (a finding) is true in this situation.

B is generally indicated as true when its sign, *A*, is true.

B is true in this situation.

For example, the following is an instantiation of the argument from sign (Hahn & Hornikx, 2016):

In this situation, there are a large number of people with digital cameras on the street.

A city, like London, is generally indicated to be a tourist destination when there are a large number of people with digital cameras on the street.

London is a tourist destination.

The scheme thus seems to be about the relationship between A and B; more specifically, about the co-occurrence between A and B. The critical questions for this argument scheme make the point about the specific relationship between A and B very clear:

CQ1: What is the strength of the correlation of the sign with the event signified?

CQ2: Are there other events that would more reliably account for the sign?

The first critical question implies that the correlation between A and B is important in determining the strength of the argument from sign. Specifically, the higher the correlation between the two event the stronger the argument. However, it is not clear from the scheme-based approach how this correlation is to be attached to the conclusion 'B is true in this situation'. There are no off-the-shelf formal tools within the scheme-based approach that would address this. The second critical questions also calls for a more graded and probabilistic approach. Namely, the critical question asks the arguer to identify other possible events that could account for the sign. However, even if one finds these events, they may plausibly vary in their likelihood. The challenge then is to reflect this likelihood onto the conclusion.

These challenges are readily addressed by the Bayesian framework (Hahn & Hornikx, 2016). The correlation between A and B can be expressed using the conditional probabilities $P(B \mid A)$ and $P(B \mid \sim A)$ where one is able to map the degree of correlation on the (likelihood) ratio of these two conditional probabilities. The likelihood ratio would then allows us to calculate the force of an argument from sign, which further enables us to determine the strength of the argument itself, i.e. the posterior probability of the conclusion.

In addition, the possibility of other events that could account for the sign and their likelihood can also be captured within the Bayesian framework. Bayesian probabilistic inference allows and can account for the interaction of multiple events (or variables). In the case of CQ2 one is required to compute the impact of other potential events (let's label them C) that can *explain* sign A on event B. To this end, one could build a Bayesian belief network model of the phenomenon called 'explaining away' (Pearl, 1988). Explaining away occurs in situations where multiple (independent) causes all compete to account for a common effect. There, knowing that the common effect holds true, further learning that one of the causes happened reduces the probability of the other causes. This reduction in the probability will depend on the prior likelihood of the causes as well as the force with which they produce there effect (explaining away and Bayesian networks are discussed in detail in Chapter 2). Explaining away seems to exactly capture the intention of CQ2. If we think of B and C as causes of a sign A, then learning there are other events C that could explain sign A would reduce the strength the conclusion B. Furthermore, one would be able to quantify the likelihood of events C as well as their impact on the strength of the conclusion B using Bayesian inference in the Bayesian belief model of the argument from sign.

This discussion of the argument from sign illustrates some of the shortcomings of the scheme-based approach that are readily addressed within the Bayesian framework. It demonstrates how one can explicate the force and the strength of an argument in probabilistic terms. Further, it shows the strength of an argument depends on the likelihood we assign to particular events, which in turn depends on the content of the argument itself. It also provides an ex-

ample of how arguments could be represented using Bayesian networks models, specifically the Bayesian network model of explaining away in the case of the argument from sign. Bayesian networks have also been used to represent other argument schemes (Hahn, Oaksford, & Harris, 2013; Hahn & Hornikx, 2016; Harris, Hahn, Madsen, & Hsu, 2016) as well as some of the 'fallacies' of argumentation (Hahn & Oaksford, 2006, 2007; Jarvstad & Hahn, 2011) other arguments schemes have also been used. Lastly, through modeling the argument from sign using the Bayesian network model of explaining away we have already seen hints of some of the relations between arguments and explanations.

1.1.2.2 Why adopt Bayesian framework as a normative standard?

We have seen that one is able to represent arguments using the Bayesian framework. However, why should one accept the Bayesian framework as a normative rational standard of argument quality? Why should one change one's beliefs regarding the strength of arguments in line with the Bayesian framework?

The Bayesian framework has its foundation in probability theory. Probability theory ensures that our beliefs are coherent with each other and that we do not end up with an inconsistent set of beliefs. Coherence, however, is not the only outcome of aligning one's beliefs with the probability theory. Namely, the so-called Dutch book theorems show that if one's degrees of belief obey the axioms of probability, then if we place bets in line with our degrees of belief, there is not a set of bets that will incur a sure loss (Ramsey, 2016; Vineberg, 2016). In other words, making sure that our beliefs align with the axioms of probability will ensure that we do not place bets against the nature that would results in a sure loss (Hahn, 2020).

Furthermore, using conditionalization to update our beliefs in light of new evidence (i.e. the premises of an argument) will uniquely minimize the inaccuracy of our beliefs across all possible worlds (that is, regardless of how the world turns out), on the condition that inaccuracy is measured with a proper scoring rule, such as the Brier score, and that those worlds are finite (see, e.g., the formal results outlined in Pettigrew, 2016). That is, the Bayesian framework specifies how we *should* change our beliefs, if we wish those beliefs to be accurate.

The rationality and the normativity of the Bayesian framework then comes from the fact that a Bayesian agent's beliefs will be coherent, never resulting in a sure loss if the agent is to place a bet, and minimally inaccurate.

1.2 Explanations and/as arguments

At the start of this chapter I suggested that explanations and arguments are similar in that they share certain features. They are both often answers to the 'why' questions and include 'because' which precedes the provision of argumentative support and the provision of an explanation, and they both constitute reasons (Hahn, 2011). For example,

> The LaTeX file won't compile, because there is a bug in my code.

could be considered as an argument for the conclusion that the LaTeX file will not compile as well as an explanation for why the file is not compiling.

It is thus often hard to distinguish between arguments and explanations. Indeed, some have even identified explanations with arguments. For instance, one of the most detailed and influential accounts of explanations is due to Carl

Hempel (Hempel & Oppenheim, 1948; Hempel, 1965). Explanations according to Hempel are logically valid arguments where premises are sentences which are "adduced to account for the phenomenon" (also called 'explanans') and the conclusion is a sentence "describing the phenomenon to be explained" (also called 'explanandum') (Hempel, 1965, p. 247). The logically valid argument from the start of this chapter could also be considered an explanation according to Hempel:

All pigs have a snout.

Snowball is a pig.

Therefore:

Snowball has a snout.

The above argument would be an explanation for why Snowball has a snout (explanandum), which is accounted by a set of sentences 'All pigs have a snout' and 'Snowball is a pig' (explanans).

The close relationship between arguments and explanations is also exemplified in non-deductive types of reasoning, such as inference to the best explanation (IBE) (Harman, 1965; Lipton, 2003). IBE plays a crucial role in both everyday and scientific reasoning contexts, with some even going as far as to suggest that IBE is *the* quintessential way of arguing for theories in science (e.g. Lipton, 2003; Psillos, 2005; Williamson, 2016). To illustrate IBE, imagine that you leave a piece of cheese on the kitchen table in the evening. The next morning, you find that the cheese is gone (except for a few crumbs), and you see that there is a small hole in the bottom of the wall. The best explanation for these observations is that a mouse visited the kitchen in the night, and you subsequently infer the truth of this hypothesis on the basis of its explanatory power

(the example is due to Van Fraassen, 1980, pp. 19–20). Similarly, Halley (1752) argued that the best explanation of the observed comets of 1531, 1607, and 1682 was that these observations were all due to a single comet (later named 'Halley's comet') that made three revolutions in an elliptical orbit around the sun with a period of 75–76 years.[1] That the one-comet hypothesis best explains the evidence raises our confidence in that hypothesis. The general idea is thus the following: explanatory considerations are truth-conducive and that a hypothesis is the best explanation is a mark of the truth of that hypothesis. The decades long debate on whether an explanation exhibiting certain explanatory considerations that other explanations do not should be considered more likely to be true is still quite present (Douven, 2013; Harman, 1967; Henderson, 2013; Lipton, 2003; Thagard, 1978). Likewise, both the theoretical and empirical work on explanatory virtues, markers of good explanations, is still ongoing (Douven & Schupbach, 2015; Lombrozo, 2007; Thagard, 1989; Wojtowicz & DeDeo, 2020).[2]

IBE is thus another example where arguments and explanation come together and are difficult to distinguish. However, despite their similarities, there are also differences between arguments and explanations. For one, the goals of arguments and explanations seem to be different. Arguments seek to increase conviction or confidence in a claim or to remove doubt about a claim that has not been university accepted. Explanations seek to increase understanding of something that has already been accepted as a fact (Antaki & Leudar, 1992; Hahn, 2011; Walton, 2004a). Let us consider again the following statement:

[1] In other words, the one-comet hypothesis is a common cause that is able to explain or *screen off* all three events (see 'The principle of the common cause' or 'conjunctive fork' in Reichenbach, 1953/1991; Salmon, 1984).

[2] Explanatory virtues are discussed in detail in Chapter 3.

The LaTeX file won't compile, because there is a bug in my code.

If the above is considered an argument, then we do not know yet whether the file will compile and the goal is to convince someone that it will not. If, on the other hand, the above statement is considered an explanation, then we already know that the file does not compile and the goal is to provide understanding for why that is the case. Antaki and Leudar (1992), however, argue that even this distinction is sometimes blurred, particularly because whether or not a statement is interpreted as an argument or as an explanation may depend on the perceived intentions and assumptions of the speaker.

1.3 Three dimensions of explanation

Arguments and explanations thus often seem similar and closely related, albeit with some potential differences. What, if anything, can we learn from this relationship between arguments and explanations? What are some of the ways that arguments and explanations interact?

The answer to these questions will inevitably depend on how someone interprets what an explanation is. The way we understand explanations seems to vary across different parts of the explanation literature. We have seen that an explanation is variously a hypothesis or a claim, or evidence that can support other claims, or an answer to a question. A conceptual map of the different notions of explanation, thus, could then be useful in helping us better place the discussion on the relationship between explanations and arguments. I provide one such map by looking at three distinctions found in the literature on explanation.

The first distinction comes from the cognitive science and psychology literature on explanations. There one can find the view that we can understand explanations either as products or as processes (see Lombrozo, 2012). From the product perspective, an explanation is a hypothesis or a claim that accounts for evidence (explanandum) when prompted to do so. All the explanations that I have introduced thus far can be considered as products. In contrast, explanations can also be viewed as a cognitive activity (process) that has as its goal to generate explanation 'products'. Here the focus is not on the very product of explanation, but rather on engaging in the particular cognitive activity of trying to explain something. The work on explanations in philosophy and in part in psychology has mostly focused on explanations as products. The empirical work in psychology has also studied explanations as processes, mostly focusing on the characterises and consequences of engaging in such a processes on, for instance, learning (for an overview see Lombrozo, 2012).

Another distinction, mostly stemming from the work in computer science on expert systems, is between explanations of outputs (evidence) and explanations of inference processes (Lacave & Díez, 2002, provide a detailed overview). Explanations of outputs or evidence relate to identifying specific factors that can account for the observed evidence. For example, an explanation of evidence in a medical context would consist of determining (a set of) diseases that best account for the symptoms, test results, etc. Explanations of inference or reasoning processes pertain to providing an account of how specific factors (inputs) lead to the observed evidence (output). More specifically, they provide accounts of which inferential steps have been taken (by an AI system) so that

the factors (input) produce the observed evidence (output).[3]

Finally, the vast majority of the work on explanation in philosophy, psychology, and partly computer science has focused on studying explanation from an intrapersonal point of view. In other words, explanations have been studied mostly in isolation and from the perspective of an individual which does not interact with others in the society. However, in both the everyday and scientific contexts, we provide explanations to someone else or we receive explanations from someone else (Hilton, 1990; Lacave & Díez, 2002). In other words, explanations also have a clear interpersonal dimension.[4]

We can summarize these three distinctions using a graphical representation. Figure 1.1 provides an illustration. The three distinctions are represented as the three dimensions of an explanation space. This conceptual representation of explanations helps us to more easily navigate the discussion regarding the relationship between arguments and explanations.

The three colored intersection points between the three explanation dimension in Figure 1.1 are points that are being addressed in this thesis. Each of the three intersection points is explored in one of the chapters of this thesis. I next describe each of the intersection points.

1.3.1 Explaining evidence

The discussion regarding the argument from sign has already suggested a potential way in which arguments and explanations relate. The Bayesian treatment of CQ2, which asked for other events (C) that would be able account for

[3]Chapter 3 provides a detailed discussion of this distinction, with a specific focus on Bayesian networks.

[4]Chapter 4 discusses in detail this interpersonal aspect of explanations.

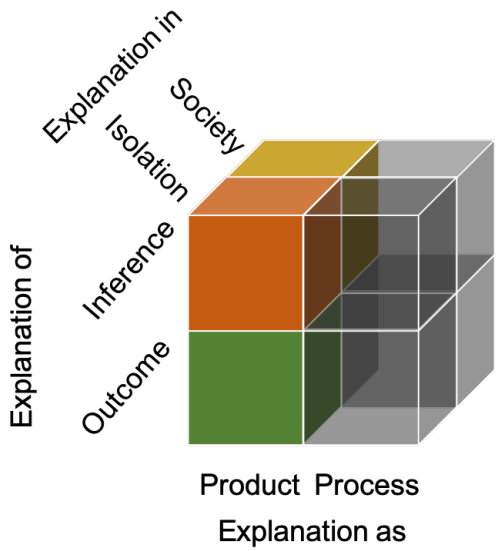


Figure 1.1: The three dimensions of explanation. The small cubes colored green, orange, and yellow are the points of intersection of these three dimensions that are discussed in this thesis.

the sign, included modeling the argument from sign using the Bayesian network model of explaining away. The intuition of CQ2 captured by this model is that learning that some other events C can account for or *explain* the presence of the sign affects the probability of the original event B. In other words, by explaining the sign using other events C, one can change the strength of the argument for B, implying that explanations can affect the strength of arguments.

We can explicate the notion of explanation in the context of the argument from sign using the three dimensions of explanation. These explanations are events or claims and hypotheses in the same manner that the original event B is a claim or a hypothesis. In that sense, these explanations are products rather than cognitive processes. Further, the explanations in the context of the argu-

ment from sign are explanations of the sign which itself is considered to be evidence (or outcome) in the Bayesian argumentation framework, making them explanations of outcomes. Lastly, the argument from sign is often considered in an intrapersonal context where the social aspects such as the relationship between the person providing an argument and the person receiving an argument is not explicitly accounted for. Therefore, the explanations employed in the argument from sign are products which explain outcomes (evidence) and are often made in an intrapersonal context. In Figure 1.1 this intersection is colored green.

The argument from sign is, however, an instantiation of an even more general pattern of causal-probabilistic reasoning, namely explaining away. Explaining away is potentially then a fruitful ground for the investigation of the relationships between arguments and explanations as specified in this section. In Chapter 2 I do exactly that: I discuss explaining away and explore the ways in which explanations have an effect on confidence in our beliefs.

1.3.2 Explaining the argument

The argument from sign not only illustrates one potential way in which explanations and arguments interact, but it also provides an illustration of how arguments can be modeled using (causal) Bayesian networks; in particular, the causal Bayesian network for explaining away can be used to model the argument from sign.

Now, causal Bayesian networks are considered experts systems (Lacave & Díez, 2002) where another aspect of explanations is clearly demonstrated. Namely, often simply finding factors (or variables in the case of Bayesian net-

works) that account for the evidence or outcomes is not sufficient. What is sometimes required is an explanation of the reasoning processes, that is an explanation of how certain factors (inputs) lead to evidence or outcomes. This is sometimes achieved by providing a chain of inference that shows how an outcomes follows from the inputs (Moulin, Irandoust, Bélanger, & Desbordes, 2002; Scott, Clancey, Davis, & Shortliffe, 1977; Walton, 2004a). To use argumentation terminology, these kinds of explanations show how the evidence impacts the hypotheses by connecting the evidence and hypotheses via inferential steps, i.e. they explain the inference in an argument.

The explanations of inference processes could still be considered as explanation products, as they are still claims or hypotheses rather then cognitive processes that people engage in to produce explanation products. Also, similarly to the notion of explanations discussed in the previous section they are often made in an intrapersonal context. In Figure 1.1 the location of these explanation in an explanation space is marked with an orange cube.

In Chapter 3 I explore explanations of reasoning processes in causal Bayesian networks. As casual Bayesian networks could also be considered tools for representing, generating and evaluating arguments, the insights regarding explanations of reasoning processes in causal Bayesian networks should translate to arguments.

1.3.3 Social explanation

In the previous section explanations of inference processes are considered in an intrapersonal context, i.e. a context where the social exchange of explanations is not present. However, argumentation is a social activity. It is present in

a discourse between two or more people (Van Eemeren et al., 2013). Often, people provide arguments for a particular claim in order to convince *others* and change beliefs of *others* about the claim. Further, it is not uncommon that some people are more and some are less convinced by the same arguments. These are all aspects of the social character of arguments and are explicitly captured by the Bayesian argumentation framework (Oaksford & Chater, 2020).

If explanations are in many aspects similar to arguments, one would expect that explanations also have a social aspect. Indeed, both the psychology and the computer science literature agree that explanations have an important social aspect (Hilton, 1990; Miller, 2019; Moulin et al., 2002; Tešić & Hahn, in press). Whether they explain inferences of an AI system or some other inference processes, providing an explanation is a communicative act in that it includes an explainer (a person or a machine providing an explanation) and an explainee (a person receiving an explanation). This notion of explanation still accounts for the inference processes and thus can be considered as an explanation product, but in contrast to the notion of explanations from the previous section, explanations viewed in this way are distinctly immersed in a social context. The location of these explanations in the 3-dimensional explanation space is marked by the yellow cube in Figure 1.1. In Chapter 4 I consider some of the effects of understanding explanations of reasoning processes in a social context on our beliefs about particular claims.

1.4 Prospectus

In this thesis I will address the three notions of explanation and their relationship to argument. In Chapter 2 I explore the effects of explaining evidence in

the context of causal-probabilistic reasoning. I investigate how providing an explanation of evidence affects the confidence people have in the claims in two causal models of explaining away. In Chapter 3 I provide further theoretical background regarding the different notions of explanations. I discuss factors, such as explanatory virtues, that make explanations 'good' explanations. I explore these factors with respect to the notion of explanations as inference processes in the context of causal Bayesian networks. In Chapter 4 I focus on one of the aspects of explanations when embedded in a social context, namely the reliability of the explainer. I explore how providing an explanation, the reliability of an explainer, and confidence in beliefs relate and affect each other. Finally, in Chapter 5 I discuss the findings of this thesis, pointing to some of the implications and potential directions for future research.

2

Argument and explanation in causal reasoning

Chapter 1 introduced the three dimensions of explanation: (1) explanations as products or processes, (2) explanations of outcome or inference, and (3) explanations in an intrapersonal or interpersonal context (see Figure 2.1). In this chapter, I discuss product explanations made in an intrapersonal context that

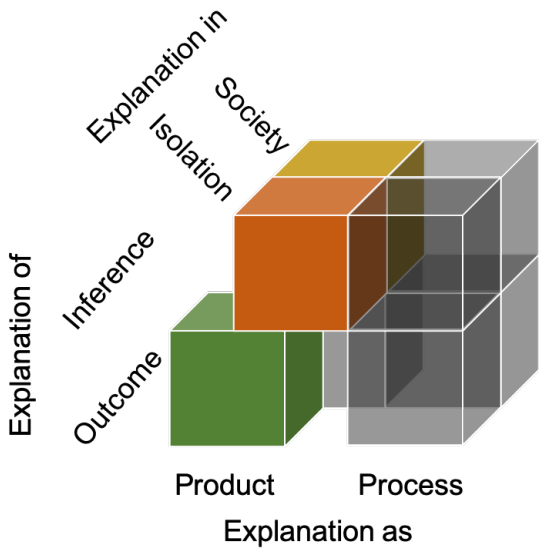


Figure 2.1: The three dimensions of explanation. This chapter discusses the intersection that corresponds to the green cube.

account for specific outcomes. More specifically, I am going to discuss causal explanations whereby effects (outcomes) are explained by appealing to causes that can account for these effects.

One of the most common, and the simplest, way of explaining something is to appeal to its cause (see Lombrozo & Vasilyeva, 2017, for a review on causal explanations). For example, a doctor may appeal to a particular class of viruses (cause) to explain a rash (effect) on a patient's body or an engineer may appeal to defects in cast iron (cause) to explain the bridge collapse (effect). However, it is not uncommon that there may be multiple causes that can account for the same effect. A rash could be caused both by a viral and a bacterial disease; a bridge may collapse both because of the defects in cast iron and because the structure of the bridge was not able to withstand an earthquake. Furthermore,

the same effect could be caused by multiple *independent* causes. For instance, asthma and flu are independent, that is, having asthma does not make one more or less likely to have flu and vice versa; but both asthma and flu can cause a cough.

The situations where multiple independent causes can account for the same effect are particularly pertinent to the question of how explanations and argument are related. This is because in these situations a pattern of causal reasoning called 'explaining away' emerges. Namely, learning that one of the independent causes happened (whilst knowing that the effect is present) will decrease the probability of (confidence in) the other cause even though the causes are (initially) probabilistically and causally independent. The fact that one of the cases has happened is sufficient to explain the effect makes the other case less likely to occur.

In Chapter 1 we have seen how the argument from sign scheme has similar characteristics to explaining away. What is more, one can even model argument from sign as an explaining away reasoning (Hahn & Hornikx, 2016). Explaining away is thus a good example of a reasoning schema where explanation and argument come together. It incorporates diagnostic reasoning—the reasoning from effect to causes, that is the assessment of the probability of a cause of after the effect is known to have happened—and intercausal reasoning where additionally learning that one cause has happened is (i) sufficient to *explain away* the effect (evidence) and (ii) it has as a consequence that the probability of (confidence in) the other cause is affected. That is, in explaining away there is a clear impact of explanations on confidence, where explanations could be thought of as a certain kind of evidence or premises in an argument. Because explanation

and argument come together in explaining away, the situations exemplifying explaining away lend themselves to studying the relationships between explanations and arguments.

The explanations in explaining away have the following simple form: a cause has happened and as such it has (partially) accounted for the effect. Here, the explanations account for a specific *outcome* (i.e. the effect). They are also *products* in a sense that these explanations can be thought of as propositions or judgements that address an explicit request for why the effect happened. Lastly, explaining away has often been studied in the context of individual and intrapersonal reasoning rather than social and interpersonal reasoning. In this chapter, I will also be looking at explaining away from the *individual* reasoning point of view. In the 3-dimensional explanation cube, the explanations in explaining away situations that I will be focusing on in this chapter are then viewed as products that account for specific outcomes and are made in an intrapersonal context (see Figure 2.1).

The chapter has two parts. In the first part I focus on explaining away in the form that is often studied in the literature: two causes and one common effect. There, I introduce explaining away as a pattern of reasoning and provide an overview of empirical work on explaining away. I further empirically test two hypothesis that aim to account for the empirical findings regarding explaining away. In the second part, I discuss an extension of explaining away to situations where multiple effects could be accounted for by the same causes. In particular, I focus on a (causal) structure where there are two causes and two common effects. Here, I discuss different ways to model these kinds of situations. The prediction of the different models are then empirically tested.

2.1 Explaining away: Accounting for a single common effect[1]

2.1.1 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. It is not rare that there may be many possible causes and effects that are interconnected in a 'causal web', which makes these judgments difficult to make (see e.g. Cruz et al., 2020).

The complexity of causal webs is not the only aspect of causal reasoning that makes it hard. Many of the causal judgment also occur *under uncertainty* and getting the causal-probabilistic judgements right is then exceedingly hard. However, having correct causal judgements is important, particularly in specialized contexts where getting them wrong can lead to deleterious consequences. 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

[1]This section is based on work from Tešić, Liefgreen, and Lagnado (2020) and Liefgreen and Tešić (in press).

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 that I have mentioned in the beginning of this chapter, i.e. explaining 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).

[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 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).

The above example of causal-probabilistic reasoning involves only two causes and one common effect; but even this causal situation with a seemingly simple causal web leads many people to erroneous reasoning. In this part of Chapter 2, I will discuss why people get the causal-probabilistic judgements in explaining away situations wrong. I start by briefly introducing Causal Bayesian Networks (CBNs), a tool for graphical representation of causal-probabilistic relations and causal-probabilistic reasoning. I will then present a CBN model for explaining away. Next, I will outline previous empirical work on explaining away in the psychological literature and point to the potential shortcomings of this work. Finally, I will discuss motivations and details of the experimental work presented in this part.

2.1.2 Causal Bayesian networks

Bayesian networks are a tool for graphical representation of probabilistic relationships among a number of variables and for making inference regarding these variables (Neapolitan, 2003; Pearl, 1988). They are directed acyclic graphs (DAGs) with nodes representing random variables[4] and arrows pointing only in one direction (hence directed) and encoding probabilistic (in)dependency relations between variables. Furthermore, in Bayesian networks there cannot be a path that, following the arrows, starts and finishes at the same node (hence acyclic). When arrows also have a causal interpretation, that is when an arrow between two nodes implies not just that one variable (say variable *B*) is proba-

[4]In this thesis, 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.

bilistically dependent on the other (say variable A), but also that B is causally dependent A (or A is a cause of B), then we say that Bayesian network is a causal one (i.e. a CBN).

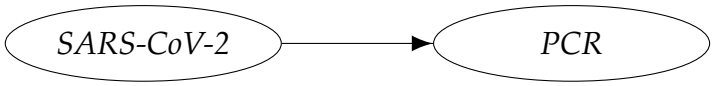


Figure 2.2: An example of CBN model.

For example, the network in Figure 2.2 is a causal Bayesian network. It has two nodes representing two random variables *SARS-CoV-2* and *PCR*, each taking two values: SARS-CoV-2 (indicating that a person does have SARS-CoV-2 virus) and ∼SARS-CoV-2 (indicating that a person does not have the virus) *and* PCR (indicating a positive PCR test result) and ∼PCR (indicating a negative PCR test result). The arrow between the two variables indicates the causal-probabilistic relationship between them: the results of a PCT test are probabilistically and causally dependent on whether someone has the virus.

In order to perform quantitative computations using a CBN, 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. In the network in Figure 2.2, the node *SARS-CoV-2* is a root and one thus needs to specify the prior probability of a person having SARS-CoV-2, that is, the probability that a person has the virus before seeing the PCR test results (formally written as $P(\text{SARS-CoV-2})$). This could be a proportion of people in the population that have the virus (which is hard to estimate), but it can also be a clinician's be-

lief that a specific person has the virus before seeing their test results.[5] As the SARS-CoV-2 variable is binary, the prior probability that a person does not have the virus is simply $1 - P(\text{SARS-CoV-2})$. To complete the parameterization of the network in Figure 2.2 one would also need to specify the probability that a person receives a positive PCR test given that they have the virus: this is known as the true positive or sensitivity rate and is formally written as $P(\text{PCR} \mid \text{SARS-CoV-2})$; and the probability that the person receives a positive PCR test given that they do not have the virus: this is known as the false positive rate rate and is formally written as $P(\text{PCR} \mid \sim\text{SARS-CoV-2})$. The probability that a person receives a negative PCR test given that they have the virus i.e. $P(\sim\text{PCR} \mid \text{SARS-CoV-2})$ (known as the false negative rate) and the probability that a person receives a negative PCR test given that they do not have the virus, i.e. $P(\sim\text{PCR} \mid \sim\text{SARS-CoV-2})$ (known as the true negative or specificity rate) are simply $1 - P(\text{PCR} \mid \text{SARS-CoV-2})$ and $1 - P(\text{PCR} \mid \sim\text{SARS-CoV-2})$ respectively.

The true positive and false positive rates describe the operating characteristics of a test and can be used to assess the strength of a test. When the ratio of $P(\text{PCR} \mid \text{SARS-CoV-2})$ and $P(\text{PCR} \mid \sim\text{SARS-CoV-2})$—also known as the likelihood ratio—is greater than 1, the higher the ratio, the stronger the evidence for the virus from a positive test. A likelihood ratio that is equal to 1 would indicate that a positive (or a negative) test is completely uninformative with regards to whether the person has the virus or not; and a positive result from a test with the likelihood ratio that is lower than 1 would make it more likely

[5] For example, Watson, Whiting, and Brush (2020) provide advice to clinicians on the variably of priors among different people and their importance in interpreting COVID-19 test results.

that a person *does not* have the virus.

After all the parameters are specified, one can perform quantitative computations. For instance, one can calculate the probability that a person has SRAS-CoV-2 after receiving a positive PCR test (known as the posterior probability of SARS-CoV-2). In our example, this would be the conditional probability that the person is infected with SARS-CoV-2 given a positive PCT test result, formally written as $P(\text{SARS-CoV-2} \mid \text{PCR})$.[6] This conditional probability would then be taken as the new prior probability that a person has the virus and used in further calculations: $P_{\text{new}}(\text{SARS-CoV-2}) := P_{\text{old}}(\text{SARS-CoV-2} \mid \text{PCR})$. The conditional probability of a person being infected with the virus given a positive test result is calculated using the Bayes' theorem:

$$P(\text{SARS-CoV-2} \mid \text{PCR}) = \frac{P(\text{SARS-CoV-2}) \times P(\text{PCR} \mid \text{SARS-CoV-2})}{P(\text{PCR})} \tag{2.1}$$

I have already introduced all terms in Bayes' theorem with the exception of $P(\text{PCR})$ which is the prior probability that a person receives a positive PCT test result. However, one can calculate that probability by using the law of total probability: $P(\text{PCR}) = P(\text{SARS-CoV-2}) \times P(\text{PCR} \mid \text{SARS-CoV-2}) + P(\sim\text{SARS-CoV-2}) \times P(\text{PCR} \mid \sim\text{SARS-CoV-2})$. Therefore, Bayes' theorem effectively combines together the prior (that is, our confidence that a person has the

[6]Note, however, that the posterior probability and the conditional probability are not always equivalent. For example, a clinician could have initially told to a patient that they have tested positive and communicated the probability of SARS-CoV-2 under this condition, i.e. $P(\text{SARS-CoV-2} \mid \text{PCR})$. At a later point in time, the clinician could realise that they have mistaken the patient's test result with someone else's and that the test results of the patient in question are still unknown. In this case, the new, posterior probability of the patient having been infected with the virus is the unconditional probability $P(\text{SARS-CoV-2})$.

virus before seeing the test results) and the likelihoods (that is, the evidential strength of a test result) to calculate the posterior probability (that is, our new updated confidence in a person having the virus following the test results).

Using Bayes' theorem to calculate the probability that a person is infected with SARS-CoV-2 after receiving a positive PCT test result is an instance of diagnostic reasoning (Meder & Mayrhofer, 2017b). In diagnostic reasoning, one is aiming to estimate the probability of a cause after learning that the effect has occurred. It is a reasoning from effects to causes. Bayesian networks, however, can also be used to calculate the probability of an effect after learning that the cause has occurred. This kind of reasoning, from causes to effects, is called predictive reasoning. In the very simple 2-node example from Figure 2.2, the probability that a person tests positive given that they have the virus is simply the likelihood used to parameterize the CBN: that is $P(\text{PCR} \mid \text{SARS-CoV-2})$.

One can, however, easily imagine more complex situations where a person gets tested multiple times and receives multiple test results. Figure 2.3 depicts a CBN model for a situations where the same person gets tested on two different occasions and receives two PCR test results. Here, after receiving a positive PCR_1 test result, one would, via diagnostic reasoning, estimate the new probability that the person is infected with SARS-CoV-2. However, the change in the probability that the person has SARS-CoV-2 would result in a change in the probability that the person would test positive (or negative) on the second PCR_2 test. This new probability of a person receiving a positive PCR_2 test is estimated via predictive reasoning and is most likely going to be different from the likelihood $P(\text{PCR}_2 \mid \text{SARS-CoV-2})$ used to parameterize the CBN.

The CBN in Figure 2.3 illustrates another fundamental concept in causal rea-

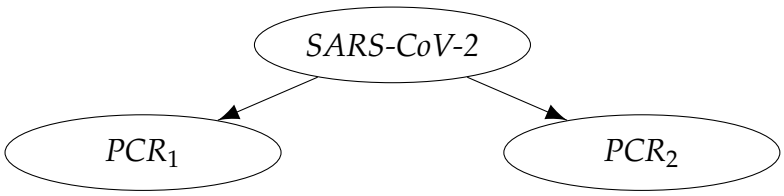


Figure 2.3: An example of a CBN model with a common cause.

soning. I have argued that learning the outcome of PCR_1 would affect the probability of the outcome of PCR_2 even though the two tests are in a sense independent; or symbolically, $P(\text{PCR}_2 \mid \text{PCR}_1) \neq P(\text{PCR}_2)$. This is because learning the outcome of PCR_1 would change the probability of the person having been infected with SARS-CoV-2, which in turn would affect the probability of the outcome of PCR_2. If, however, we first learned that the person has definitely been infected with SARS-CoV-2, then additionally learning that PCR_1 test result was positive would not change the probability of the outcome of PCR_2 test; or symbolically, $P(\text{PCR}_2 \mid \text{SARS-Cov-2}, \text{PCR}_1) = P(\text{PCR}_2 \mid \text{SARS-CoV-2})$. This is because additionally learning that the person received a positive PCR_1 test result cannot change the probability that the person is infected with SARS-CoV-2 (since we already know that they have definitely been infected with the virus), which then in turn would not change the probability of the outcome of PCR_2 test. In other words, knowing the state of a common cause makes the effects conditionally independent.

The following often mentioned example can help further illustrate the notion of conditional independence. Imagine that data analysts have found that ice cream sales and shark attacks are positively correlated: more shark attacks correspond to more ice creams sold. However, the data analysts have also found that shark attacks and ice cream sales have a common cause, namely

temperature. The higher the temperature the more people crave the cold snack increasing its sales and higher temperatures mean more people will be visiting beaches which implies more opportunities for shark attacks. Therefore, the temperature is able to explain the correlation between ice cream sales and shark attacks and knowing the temperature will result in ice cream sales and shark attacks being conditionally independent; in other words, controlling for temperature will result in ice cream sales and shark attacks being uncorrelated.

Two variables are not only conditionally independent when they share a common cause. They are also conditionally independent if they are arranged in a causal chain and there is a third variable between them. Knowing the state of that third variable would 'block' the impact of the first variable on the second one the same way knowing that a person has been infected with SARS-CoV-2 block the impact of learning the outcome of PCR_1 on the outcome of PCR_2. For example, the CBN in Figure 2.4 models a causal chain situation where the outcome of a PCR test can cause a person to self-isolate, which in turn can cause them to work from home. In this situation, if we do not know whether the person is self-isolating or not, but we do know that they have tested negative for the virus, the fact that they have tested negative will impact the probability of them working from home, as knowing that they have tested negative would presumably change the probability of them self-isolating, which in turn would change the probability that they are working from home; or $P(\text{Work from home} \mid \sim\text{PCR}) \neq P(\text{Work from home})$. However, if we know that the person is self-isolating, additionally learning of their PCR test result will not change the probability that they are working from home, or $P(\text{Work from home} \mid \text{Self-isolation}, \sim\text{PCR}) = P(\text{Work from home} \mid$

Self-isolation). This is because we already know with probability 1 that the person is self-isolating and additionally knowing the result of their PCR test will not change that probability, which means that the probability that the person is working from home will be unchanged by additionally learning the result of a PCR test.[7] We then say that the variable *Working from home* is conditionally independent from the variable *PCR* given we know the outcome of the variable *Self-isolation*.

Figure 2.4: An example of a CBN model of a causal chain.

The notion of conditional independence in CBNs is not specific to particular causal structures. Rather, it is captured by a more general principle called 'the Parental Markov Condition' (PMC) that applies to all CBNs. The PMC can be formulated as follows: a variable X in a CBN is conditionally independent from its non-descendants given its parents. Formally, this implies that $P(X \mid Parents, Non\text{-}descendants) = P(X \mid Parents)$. A node in CBN is a parent of a child node if there is an arrow going from the former to the latter. If there is a chain of nodes in a CBN, then a node that appears earlier in the chain is an ancestor of a node that appears later in the chain, and a node is a descendant of

[7] It is, of course, possible that we learn evidence with the probability that is less than 1, i.e. that we learn uncertain evidence. For example, we may not be fully sure whether a person is in self-isolation, but we have been told by multiple sources that they are, which makes us 90% confident that the person actually is in self-isolation. For how to update probabilities with uncertain evidence see Korb and Nicholson (2010).

another node if it appears later in the chain than the other node. For instance, in the CBN in Figure 2.4 the PCR is a parent node of *Self-isolation* and *Self-isolation* is a parent node of *Work from home*. PCR, however, is not a parent node of *Work from home*, but is its ancestor and *Work from home* is not a child node of PCR, but is its descendant. In the CBN in Figure 2.3, PCR_1 and PCR_2 are child and descendant nodes of their parent node *SARS-CoV-2* and PCR_1 is neither an ancestor nor a descendant of PCR_2. In the CBN in Figure 2.4, the only variable that has non-descendants is *Work from home*, namely PCR. Therefore, the PMC implies only one conditional independence relation for that network, i.e. $P(\textit{Work from home} \mid \textit{Self-isolation}, PCR) = P(\textit{Work from home} \mid \textit{Self-isolation})$.[8] In the CBN in Figure 2.3, two variables have non-descendants: the non-descendant of PCR_1 is PCR_2 and the non-descendant of PCR_2 is PCR_1. This means that the PMC implies two conditional independence relations for that CBN: $P(PCR_1 \mid \textit{SARS-CoV-2}, PCR_2) = P(PCR_1 \mid \textit{SARS-CoV-2})$ and $P(PCR_2 \mid \textit{SARS-CoV-2}, PCR_1) = P(PCR_2 \mid \textit{SARS-CoV-2})$.

Identifying conditional independence in a CBN using the PMC makes the computations of the joint probability distribution a lot less cumbersome. For instance, image a person that is tested three times for the SARS-CoV-2 virus. The CBN corresponding to this situations is presented in Figure 2.5.

[8]Note here the conditional independence is over variables not variables states. This is because the conditional independence relations hold for any combination of outcome of the three variables. Further, intuitively it seems that the reverse also holds, that is if we know that the person is in self-isolation, additionally learning that they are working from home will not affect the probability of them testing positive (or negative) for the virus. This is also true, but it is not captured by the PMC. To identify all conditional independence relations in a CBN, one can use the method called *d*-separation. For more details see Neapolitan (2003).

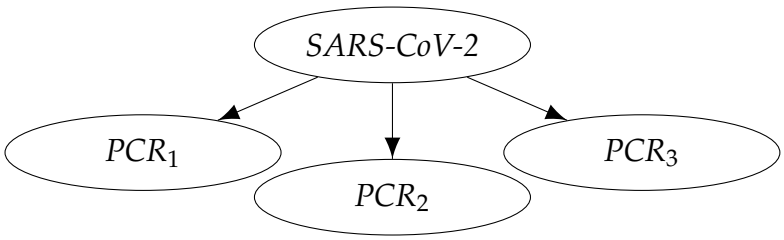


Figure 2.5: An example of a CBN model with a common cause and three effects.

To calculate the join probability distribution of all four variables, i.e. to calculate $P(\textit{SARS-CoV-2}, PCR_1, PCR_2, PCR_3)$, one could use the chain rule from the probability theory. This rule implies that $P(\textit{SARS-CoV-2}, PCR_1, PCR_2, PCR_3) = P(PCR_3 \mid PCR_2, PCR_1, \textit{SARS-CoV-2}) \times P(PCR_2 \mid PCR_1, \textit{SARS-CoV-2}) \times P(PCR_1 \mid \textit{SARS-CoV-2}) \times P(\textit{SARS-CoV-2})$. Estimating some of these conditional probabilities is notoriously hard, which makes the estimate of the joint probability distribution less reliable. The PMC significantly simplifies the estimation of these conditional probabilities as it implies that the three PCR test are conditionally independent of each other given their parent node *SARS-CoV-2*. This means that $P(\textit{SARS-CoV-2}, PCR_1, PCR_2, PCR_3) = P(PCR_3 \mid \textit{SARS-CoV-2}) \times P(PCR_2 \mid \textit{SARS-CoV-2}) \times P(PCR_1 \mid \textit{SARS-CoV-2}) \times P(\textit{SARS-CoV-2})$. Estimating these conditional probabilities is significantly easier, which makes the final estimate of the joint probability distribution more reliable. We can imagine situations where a CBN model includes a dozen or more variables where the application of the chain rule to calculate the joint probability distribution would leave us with conditional probabilities whose estimations would be practically impossible. The PMC can drastically reduce the complexity of these conditional probabilities and make their estimation significantly less difficult.

The CBNs thus have a number of features that make them quite appealing: (i) one can graphically represent the causal situations one is is trying to reason about by assigning a node to each variable/event and drawing arrows between the nodes to illustrate the causal relations between the variables/events; (ii) one can visually read off the conditional independence relations that hold among the variables/events; and (iii) the PMC and other methods for identifying conditional independences among the variables enable one to estimate the conditional probabilities with less effort then one otherwise would, particularly in more complex causal situations.

I will conclude this section on CBNs with a brief discussion of the normative aspect of CBNs. All probability distributions in CBNs are consistent with the axioms of classical probability. Therefore, the person reasoning using a CBN model cannot be Dutch booked. Furthermore, to update the probabilities the CBNs employ Bayes' theorem (and other theorems from probability theory), which implies that a person updating their beliefs using a CBN model would be minimizing their inaccuracy as discussed in Chapter 1. These features of CBN models provide us with a normative aspect that some other modeling strategies may lack. It then follows that the predictions from a CBN can be considered normative, particularly in cases where it is clear how one would model a particular causal situation.

2.1.3 Explaining away: normative account

In the previous section I have discussed two types of reasoning that feature in CBNs, the diagnostic and predictive reasoning. The third main type of reasoning that also occurs in CBNs is intercausal reasoning.

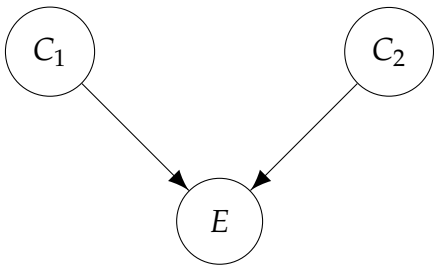


Figure 2.6: A CBN model of explaining away

Consider the graph in Figure 2.6, 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 $P(\mathrm{E} \mid {\sim}\mathrm{C}_1, {\sim}\mathrm{C}_2)$. Once one specifies these parameters, one can not only compute the probability of a cause given the effect, e.g. $P(\mathrm{C}_1 \mid \mathrm{E})$ (diagnostic reasoning) and the probability of an effect given one cause, e.g. $P(\mathrm{E} \mid \mathrm{C}_2)$ (predictive reasoning), but also the probability of one cause given the other cause, e.g. $P(\mathrm{C}_1 \mid \mathrm{C}_2)$ and the probability of one cause given the other cause and the effect, e.g. $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$. The estimation of the latter two probabilities constitutes instances of intercausal reasoning, that is reasoning from one cause to another.

As with any CBN, we can apply the PMC to the CBN in Figure 2.6. Two variables have non-descendants in this CBN: C_2 is a non-descendant of C_1 and

C_1 is a non-descendants of C_2. However, unlike in the CBNs that we have discussed in the previous section, C_1 and C_2 have no parents. By the PMC, this means that C_1 is independent from C_2 given an empty set and that C_2 is independent from C_1 also given an empty set; that is, C_1 and C_2 are *un*conditionally independent. In other words, 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).

So far, the common-effect CBN seems to align with what been said regarding conditional independence. The common-effect CBNs, however, include the following feature specific to them that makes intercausal reasoning particularly interesting. Namely, depending on the network parameterization, the two causes in a common-effect CBN that are, per the PMC, initially unconditionally independent may become conditionally dependent if we learn that the effect has happened; that is, 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. This is in contrast to the cases of conditional independence implied by the PMC in the CBNs in the previous section where two variables are conditionally independent regardless of how the network is parameterized and they do not become conditionally dependent because we learned that a variable has taken a particular state.

This brings us to explaining away. In Section 2.1.1 I have briefly introduced explaining away where I mentioned that it is a pattern of causal reasoning that occurs in situations where multiple causes compete to account for an effect. I have also pointed out that in explaining away, after we learn that an effect

has occurred, additionally learning that one of the causes has happened is sufficient to explain the effect, which would in turn affect the probability of the other cause, and more specifically it will reduce the probability of the other cause. Explaining away, thus, is an instance of intercausal reasoning and it can be modeled using common-effect CBN models such as the one in Figure 2.6 (see Pearl, 1988, 2009). For instance, we could model the example of explaining away involving physical abuse, haemophilia and bruises from Section 2.1.1 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. 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):

$$P(\mathrm{E} \mid \mathrm{C_i}, \mathrm{C_j})\, P(\mathrm{E} \mid \sim\mathrm{C_i}, \sim\mathrm{C_j}) < P(\mathrm{E} \mid \mathrm{C_i}, \sim\mathrm{C_j})\, P(\mathrm{E} \mid \sim\mathrm{C_i}, \mathrm{C_j}) \tag{2.2}$$

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 (2.2) it follows (see Morris & Larrick, 1995; Griffiths, 2001):

$$P(C_i \mid E, C_j) < P(C_i \mid E) < P(C_i \mid E, \sim C_j) \tag{2.3}$$

The inequalities in (2.3) accord with the general intuition of explaining away mentioned above and I take them as a definition of explaining away for the empirical research outlined in the this chapter (see also Rehder & Waldmann, 2017; Rottman & Hastie, 2016).

Before concluding this section and continuing onto the review of the empirical work regarding explaining away, I would like to make the following observations. 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. $P(E \mid C_i) > P(E)$) (Cheng, 1997). This is true, meaning that Inequality (2.2) (and hence the inequalities in (2.3)) holds if the causes are generative. However, it is also the case that Inequality (2.2) 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 $P(E \mid C_i) < P(E)$.[9] For example, sneezing can be prevented by taking antihistamine drugs and/or by turning on an air filtration system. Learning that

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

a person is sneezing will decrease the probability 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 (2.2) holds with regards to the generative/inhibitory nature of causes are presented in Appendix A.1.

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

2.1.4 Explaining away: empirical account

Despite the ubiquity and importance of explaining away in a wide range of contexts, including social attribution, medical diagnosis and legal domains (Kelley, 1973; Pearl, 1988; Rottman & Hastie, 2016), empirical research on explaining away in the psychological sciences adopting the constrained definition outlined by the inequalities in (2.3) is somewhat limited and has insofar yielded mixed findings (for an overview see Rottman & Hastie, 2014). Over-

all, however, it appears that human explaining away inference, even in simple three-node common-effect causal structures (see Figure 2.6), 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, meaning that after learning that one cause has happened people underadjust the probability of the other cause (Davis & Rehder, 2017; Fernbach & Rehder, 2013; Morris & Larrick, 1995; Rehder & Waldmann, 2017; Rottman & Hastie, 2016; Sussman & Oppenheimer, 2011); in some cases, the studies have recorded a behaviour directly opposite to that of explaining away: $P(C_i \mid E, C_j) > P(C_i \mid E, \sim C_j)$ (Fernbach & Rehder, 2013; Rehder, 2014a) or $P(C_i \mid E, C_j) > P(C_i \mid E)$ (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. By contrast, 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 proba-

bilities 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.

2.1.5 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 are commonly found in these studies.

2.1.5.1 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 $\Delta_1 = P(C_i \mid E) - P(C_i \mid E, C_j)$ and $\Delta_2 = P(C_i \mid E, {\sim}C_j) - P(C_i \mid E)$ (see Figure 2.7). 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 this chapter I 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

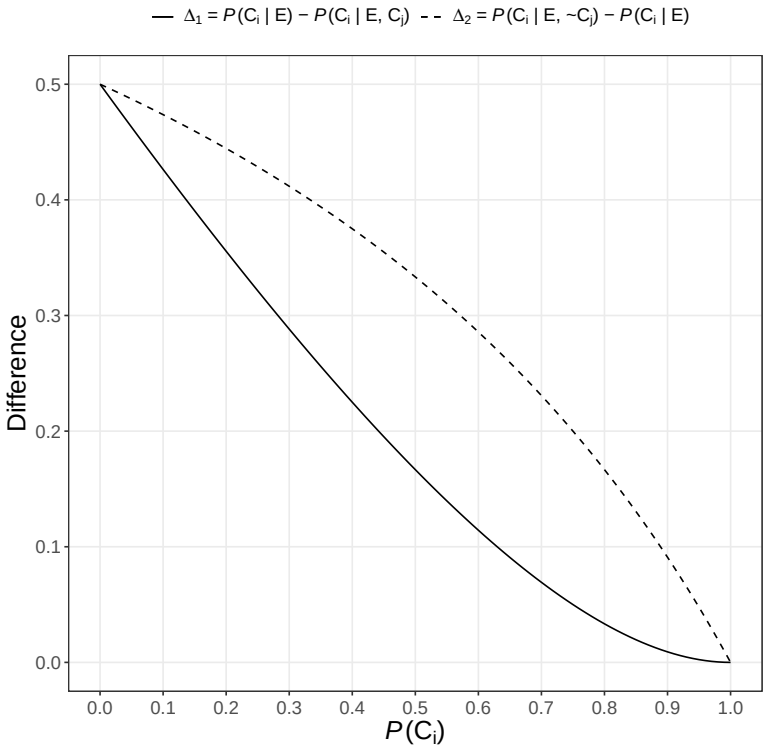


Figure 2.7: The difference $\Delta_1 = P(\mathrm{C_i} \mid \mathrm{E}) - P(\mathrm{C_i} \mid \mathrm{E}, \mathrm{C_j})$ and $\Delta_2 = P(\mathrm{C_i} \mid \mathrm{E}, \sim\mathrm{C_j}) - P(\mathrm{C_i} \mid \mathrm{E})$ as a function of the priors ($P(\mathrm{C_i})$). The prior probabilities of the causes are assumed to be equal in this figure. Further, the figure assumes deterministic set-up, i.e., $P(\mathrm{E} \mid \mathrm{C_1}, \mathrm{C_2}) = P(\mathrm{E} \mid \mathrm{C_i}, \sim\mathrm{C_j}) = 1$ (where $i, j \in \{1, 2\}$), and $P(\mathrm{E} \mid \sim\mathrm{C_1}, \sim\mathrm{C_2}) = 0$.

in the model to vary the normative amount of explaining away.

2.1.5.2 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. $P(C_i \mid C_j) \neq P(C_i \mid \sim C_j)$ (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 understanding 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 $P(C_i \mid C_j) = .45$ and $P(C_i \mid \sim C_j) = .35$ (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. $P(C_i \mid C_j) = .25$ and $P(C_i \mid \sim C_j) = .25$), one gets that $P(C_i \mid E) = .54$ and $P(C_i \mid E, C_j) = .55$ (see Appendix A.2). So, given the participants' reported positive correlation between the causes, the difference between $P(C_i \mid E)$ and $P(C_i \mid E, C_j)$ is now negligible and slightly goes in the opposite direction to explaining away. Furthermore, these new normative probability values for $P(C_i \mid E)$ and $P(C_i \mid E, C_j)$ closely approximate average participants' estimates: $P(C_i \mid E) = .58$ and $P(C_i \mid E, C_j) = .56$ (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 the studies below I 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.

2.1.5.3 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 (2.3). It is then important to empirically explore whether people understand this relational na-

ture 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. $P(C_i \mid E, C_j)$, 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. $P(C_i \mid E)$.

Despite the intuitive importance of asking qualitative relational questions when testing for explaining away, to the best of my knowledge only few studies have employed such or similar methods: Ali, Chater, and Oaksford (2011, Experiment 2), Hall, Ali, Chater, and Oaksford (2016), Liefgreen et al. (2018), and Rehder (2014a). The research presented here builds on these studies and complements quantitative questions asking for numerical probability estimates of, for example, $P(C_i \mid E, C_j)$, with qualitative relational questions asking them to consider whether $P(C_i \mid E, C_j)$ is less than, greater than, or equal to $P(C_i \mid E)$. Further, I distinguish between *direct* explaining away which corresponds to what is usually referred to as an explaining away question, namely a question about $P(C_i \mid E, C_j)$, of course in relation to $P(C_i \mid E)$ (see for example Morris & Larrick, 1995) and explaining away as a *relational* concept captured by inequalities in (2.3) which includes the question about $P(C_i \mid E, C_j)$, but also about $P(C_i \mid E)$ and $P(C_i \mid E, \sim C_j)$ (see for example Rehder & Waldmann, 2017). This will allow for a more comprehensive view of explaining away.

2.1.6 Motivations

Due to the potential methodological confounds mentioned above and the mixed findings of the extant empirical work on explaining away, together with colleges I 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. $P(C_i \mid E)$), split the probability space between the two causes and provided answers such that $P(C_1 \mid E) + P(C_2 \mid E) = 1$ 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: $P(C_i) = P(C_i \mid E) = P(C_i \mid E, C_j)$. The explanations of the participants in cluster (2) led me to hypothesize that these participants may be interpreting probabilities as, what is called in the philosophical literature, 'propensities'.

The two conjectures regarding the two clusters prompted the current study in which I not only aimed to address the limitations of previous studies by employing a novel experimental design (see Methods section), but I have also attempted to test (i) whether people employ a strategy that I 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. I will now describe the two hypotheses in more detail and outline how I will empirically address them.

2.1.6.1 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 $P(C_1 \mid E)$ and $P(C_2 \mid E)$ added up to 1. This was particularly striking in the condition in which the stated prior probabilities were low, $P(C_1) = .2$ and $P(C_2) = .1$. In this condition, a number of participants either said $P(C_1 \mid E) = P(C_2 \mid E) = .5$ or provided a more sophisticated answer to reflect the $2 : 1$ ratio of the priors, i.e. $P(C_1 \mid E) = .67$ and $P(C_2 \mid E) = .33$ (the normative answers were $P(C_1 \mid E) = .71$ and $P(C_2 \mid E) = .36$). Participants' verbal reasoning explanations regarding $P(C_i \mid E)$ 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.[10] 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 leads to a hypothesis that when engaging in diagnostic reasoning in cases where the two (or more) independent causes become exhaustive upon learning evidence, i.e. $P(C_1 \vee C_2 \vee \ldots \vee C_n \mid E) = 1$ since $P(E \mid \sim C_1, \sim C_2, \ldots, \sim C_n) = 0$, but crucially they *do not* become mutually exclusive, i.e. $P(C_1, C_2, \ldots, C_n \mid E) \neq 0$ since $P(E \mid C_1, C_2, \ldots, C_n) > 0$, some people simply split the probability space be-

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

tween the two causes and assign each cause a .5 probability *when the causes had equal priors*. I 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 diagnostic split strategy.[11] 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 $P(\mathrm{C}_1 \mid \mathrm{E}) \approx \frac{a}{a+b}$ and $P(\mathrm{C}_2 \mid \mathrm{E}) \approx \frac{b}{a+b}$ which follows the diagnostic split predictions (see Figure 2.8).[12] As such, the diagnostic split hypothesis has some normative underpinnings and could be understood as an extreme approximation of the normative diagnostic reasoning.

Other empirical studies seem also to have found trends corresponding to the diagnostic 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

[11] I thank Ben Rottman for pointing this out to us.

[12] The fact that lower priors lead to closer to normative estimates in diagnostic split reasoning is interesting from a broader psychology of reasoning perspective. For instance, Oaksford and Chater (1994) have argued that once we assume the *rarity assumption*, namely that the probability of an antecedent and a consequent in a conditional are low, people's responses to the Wason selection task have a strong rational basis. Furthermore, the data from studies on causal reasoning seems to suggests that people often assume the rarity assumption in the case of the priors of causes. Morris and Larrick (1995, Experiment 1) found that the participants' average adopted prior on an explaining away task was .23. Similarly, J. R. Anderson (1990) derived the prior of .25 for a cause from Schustack and Sternberg (1981)'s data on causal inference with one cause and one effect. The diagnostic split reasoning, thus, adds to this work and points to the importance of rarity in (causal) probabilistic reasoning that may lead to the normative answers.

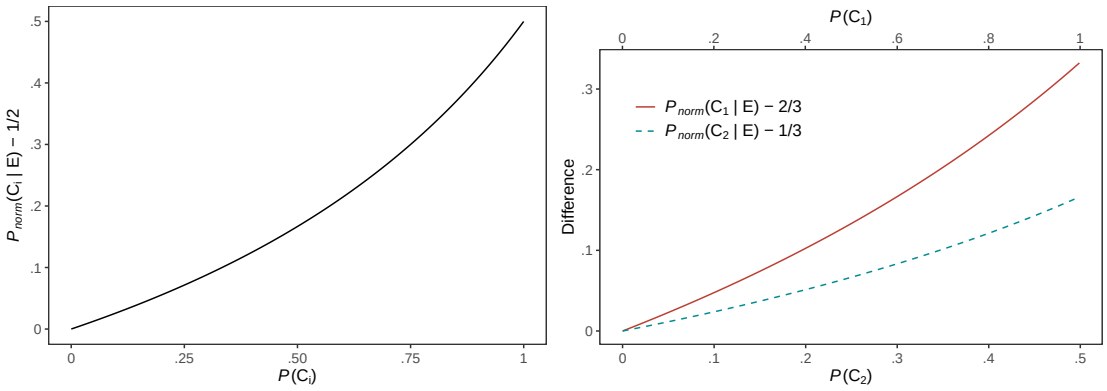


Figure 2.8: Left: the difference between the normative diagnostic reasoning ($P_{norm}(C_i \mid E)$) and the constant diagnostic split prediction of $1/2$ in the case of equal priors. Right: the difference between the normative diagnostic reasoning ($P_{norm}(C_1 \mid E)$ and $P_{norm}(C_2 \mid E)$) 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., $P(E \mid C_1, C_2) = P(E \mid C_i, {\sim}C_j) = 1$, and $P(E \mid {\sim}C_1, {\sim}C_2) = 0$. We can see that as priors are getting closer to 0 the diagnostic split hypothesis is better approximating the normative diagnostic reasoning.

both Experiment 1a where the priors were $P(C_1) = P(C_2) = .5$ and Experiment 1b where $P(C_1) = P(C_2) = .25$, 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 $P(C_1) = P(C_2) = .5$, suggests their findings fit predictions of the diagnostic split hypothesis that $P(C_i \mid E) = .5$, 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 $P(C_i \mid E) = .5$ for any $P(C_i)$.

I directly tested the diagnostic split hypothesis. In addition to low and medium priors conditions where I expected to replicate our previous findings

(i.e. I expect to find $P(C_i \mid E) = .5 \geq P(C_i)$, for $P(C_1) = P(C_2) \leq .5$), in Experiment 1 I also introduced a high priors condition ($P(C_i) > .5$). In this condition, according to the diagnostic split hypothesis, I 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, I expected to find that a number of participants will erroneously say that $P(C_i \mid E) = .5 < P(C_i)$ for $P(C_1) = P(C_2) > .5$ even though the causes are maximally strong (i.e. their strengths are 1, see Cheng, 1997).

2.1.6.2 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, $P(C_i) = P(C_i \mid E) = P(C_i \mid E, C_j)$ in both medium and low priors conditions. Through inspection of the data, I 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 to a hypothesis 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).[13] 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 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).[14] 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

[13] 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).

[14] For the purposes of this chapter I am confining myself to what Gillies (2000b) refers to as 'single-case propensity theories' (see for instance Giere, 1973).

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 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 $P(\mathrm{C}_1 \mid \mathrm{E}) = P(\mathrm{C}_1) = \frac{1}{5}$.[15] 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 in stark contrast with the normative account where these three probabilities are in general not equal.

However, like the diagnostic split hypothesis, the propensity interpretation

[15] 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' paradox' in the literature.

has its normative underpinning in the limit. Figure 2.9 shows that as the priors converge to 1, the normative diagnostic reasoning estimates approach the predictions of the propensity interpretation, i.e. that $P(C_i \mid E) - P(C_i) = 0$. Furthermore, when the explaining away set-up is deterministic (as in the experiments in this chapter), then even normatively it holds true that $P(C_i \mid E, C_j) = P(C_i)$. 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 2.9 we can see that the propensity interpretation approximates normative diagnostic reasoning within .1 error when the priors are higher than .63. From Figure 2.8 on the left we can see that the diagnostic split hypothesis approximate 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.

I thus hypothesise that the propensity interpretation, which predicts that $P(C_i \mid E, C_j) = P(C_i \mid E) = P(C_i)$[16], could be partly driving the insufficiency

[16]In general, the propensity interpretation would also predict that $P(C_i) = P(C_i \mid C_j) = P(C_i \mid \sim C_j) = P(C_i \mid \sim E) = P(C_i \mid C_j, \sim E) = P(C_i \mid \sim C_j, E) = P(C_i \mid \sim C_j, \sim E)$. However, given that in this chapter I have adopted a deterministic set-up, it is not possible for $P(C_i \mid \sim E)$, $P(C_i \mid C_j, \sim E)$, and $P(C_i \mid \sim C_j, \sim E)$ to equal $P(C_i)$ since if we learn that evidence does not obtain that means that both causes are false with probability 1. Furthermore, I did not predict

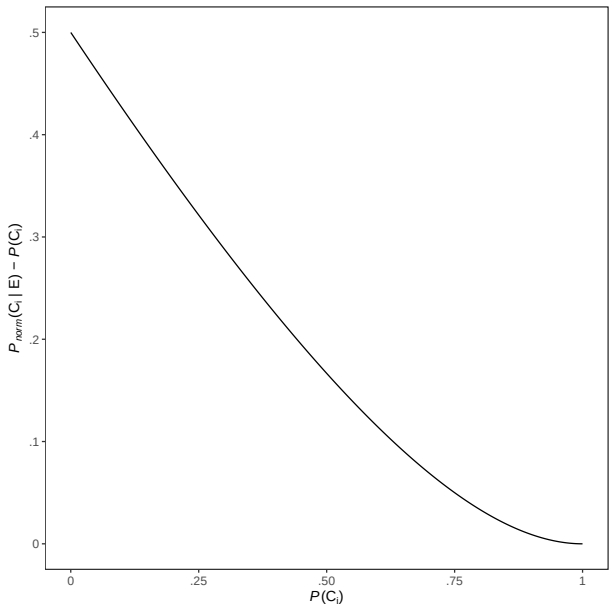


Figure 2.9: The difference between the normative diagnostic reasoning ($P_{norm}(\mathrm{C_i} \mid \mathrm{E})$) and the prior probability of the causes in the case of equal priors. The figure assumes deterministic set-up, i.e., $P(\mathrm{E} \mid \mathrm{C}_1, \mathrm{C}_2) = P(\mathrm{E} \mid \mathrm{C_i}, {\sim}\mathrm{C_j}) = 1$ (where $i, j \in \{1, 2\}$), 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.

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 $P(C_1 \mid E)$ is as equally likely as $P(C_1 \mid E, C_2)$. 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).[17]

that $P(C_i \mid \sim C_j, E)$ would be accounted by the propensity hypothesis as, in the deterministic set-up, it becomes a simple logic inference (see below) as $P(C_i \mid \sim C_j, E) = 1$. Lastly, although the propensity hypothesis predicts that $P(C_i) = P(C_i \mid C_j) = P(C_i \mid \sim C_j)$ I did not focus on these probability estimates when it came to the propensity interpretation (however, see the results sections regarding the independence of causes) as these results are equality predicated by the normative account.

[17] 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 be particularly salient when evidence is symmetrically balanced, which could be expounded as cases where $P(C_1 \mid E) = P(C_2 \mid E) = \ldots = P(C_n \mid E)$. These cases seem to correspond to cases in diagnostic reasoning where partici-

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

pants 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), I 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.

> event. This concept does not always coincide with the lay interpretation of probability. People 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.

2.1.7 Experiment 1[18]

The aim of this experiment was two-fold: (i) to test people's intuitions in explaining away contexts and (ii) to 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 I used a novel experimental design

[18]This experiment was conducted together with Alice Liefgreen and David Lagnado (Tešić et al., 2020).

that not only addressed previously mentioned methodological confounds of previous studies, but additionally allowed for a manipulation of two main factors: the prior probabilities of causes and the properties of cover stories within which the same common-effect three node structure was embedded.

2.1.7.1 Manipulations

Prior probabilities of causes By manipulating priors of causes I 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, I created three conditions in which the prior probabilities of the causes were either low—$P(C_1) = P(C_2) = .2$—medium—$P(C_1) = P(C_2) = .5$—or high—$P(C_1) = P(C_2) = .7$. In all conditions, the presence of at least one cause entailed the presence of the effect: $P(E \mid C_1, C_2) = P(E \mid C_i, {\sim}C_j) = 1$ (where $i, j \in \{1, 2\}$); and absence of both causes entailed absence of the effect: $P(E \mid {\sim}C_1, {\sim}C_2) = 0$. The deterministic relations between the causes and the effect have as a consequence maximal normative explaining away (for a given prior probability) since $P(C_i \mid E, C_j)$ is equal to the prior probability (i.e. to $P(C_i)$). Additionally, I 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 $P(C_i \mid E)$, $i \in \{1, 2\}$, to $P(C_1 \mid E, C_2)$ is .36 and the probability change from $P(C_1 \mid E, C_2)$ to $P(C_1 \mid E, {\sim}C_2)$ 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 me 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 me to test the diagnostic split hypothesis. I 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 $P(C_i \mid E)$; for participants reasoning with medium priors I expected them to stay at .5 for both causes in $P(C_i \mid E)$; and I expected participants reasoning with high priors to lower the probabilities of causes from priors to .5 in $P(C_i \mid E)$.

Properties of cover stories In addition to manipulating prior probabilities of causes, I manipulated the properties of the cover stories. In the present study I 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 the cover stories I followed (i) the philosophy of probability literature and (ii) the general idea outlined Section 2.1.6.2 on propensity interpretation that propensities are associated with tendencies of a *physical* system that describes a par-

ticular chance set-up and that propensities are often tied with causal (or even causal-mechanistic) relationships. This would then imply that I 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 physical 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 I 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 2.1.7.2 below) that are embedded in two physical systems with clear causal-mechanistic relationships to the light bulb I argue that the propensity interpretation will be strongly pronounced in this scenario.

The second cover story 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 I 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 the study, I prompted participants to answer diagnostic reasoning and explaining away questions pertaining to the *balls* and not to the random sampling mechanism, I 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 I asked in the questionnaire, namely coins.

Finally, I 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, I argue that in this scenario the propensity interpretation is the least pronounced (if at all).

Given the above rationale, I 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.

2.1.7.2 Methods

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) $\times$ 3 (Priors condition: low, medium or high) = 9 groups ($N_{\mathrm{COINS\_LOW}} = 49$, $N_{\mathrm{COINS\_MED}} = 50$, $N_{\mathrm{COINS\_HIGH}} = 50$, $N_{\mathrm{BALL\_CONTAINERS\_LOW}} = 51$, $N_{\mathrm{BALL\_CONTAINERS\_MED}} = 52$, $N_{\mathrm{BALL\_CONTAINERS\_HIGH}} = 52$, $N_{\mathrm{DINNER\_LOW}} = 50$, $N_{\mathrm{DINNER\_MED}} = 50$, $N_{\mathrm{DINNER\_HIGH}} = 49$).

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 2.1. 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 2.1.7.2, participants in each group were required to reason with different cover stories within which I 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 (see the top image Figure 2.10). 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 (see the middle image Figure 2.10). Finally, the remaining three

groups ($\text{Group}_{\text{DINNER\_LOW}}$, $\text{Group}_{\text{DINNER\_MED}}$, $\text{Group}_{\text{DINNER\_HIGH}}$) 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') (see the bottom image Figure 2.10). For full materials visit Open Science Framework, https://osf.io/aqjkp/.

Table 2.1: Inference types and questions found in the questionnaire for Experiment 1.

Quest. Num.	Inference Type	Key Inferences	Quest. Type
1	**Priors**	$P(C_1)$	Quantitative
2		$P(C_2)$	Quantitative
3	**Independence**	$P(C_2 \mid C_1)$	Qualitative
4		$P(C_1 \mid \sim C_2)$	Qualitative
5 , 6	**Diag. Reasoning**	$P(C_1 \mid E)$-*R*-$P(C_1)$	Qual. & Quant.
7 , 8		$P(C_2 \mid E)$-*R*-$P(C_2)$	Qual. & Quant.
9 , 10	**Explaining Away**	$P(C_1 \mid E, C_2)$-*R*-$P(C_1 \mid E)$	Qual. & Quant.
11 , 12	**Logic**[19]	$P(C_1 \mid E, \sim C_2)$-*R*-$P(C_1 \mid E)$	Qual. & Quant.

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

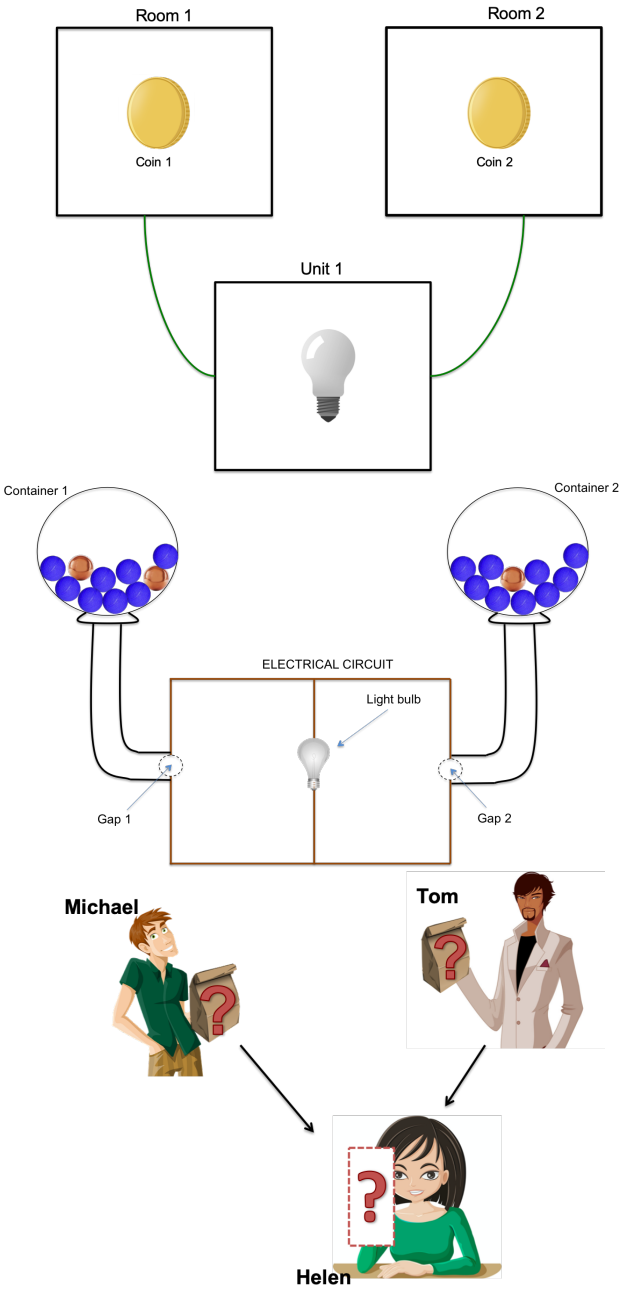


Figure 2.10: Graphical representations of the cover stories presented to participants in Experiment 1. The top image was featured in the coins cover story, the middle one in the balls and container cover story, and the bottom one in the dinner party cover story.

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 in 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).[20] 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, see Figure 2.10). 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 2.1). 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.

[19] I have labeled questions 11 and 12 as 'logic' questions, since the 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. $P(\mathrm{C}_1 \mid \mathrm{E}, \sim\mathrm{C}_2) = 1$.

[20] 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%.

Questions marked as quantitative in Table 2.1 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. $P(C_2 \mid C_1)$ given no knowledge of whether E is present or not. To investigate participants' diagnostic and explaining away reasoning I 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 enables one 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).

2.1.7.3 Results

Participants' answers to all qualitative questions in the inference questionnaire are represented in Figure 2.11 and their responses to all quantitative questions are in Figure 2.12.

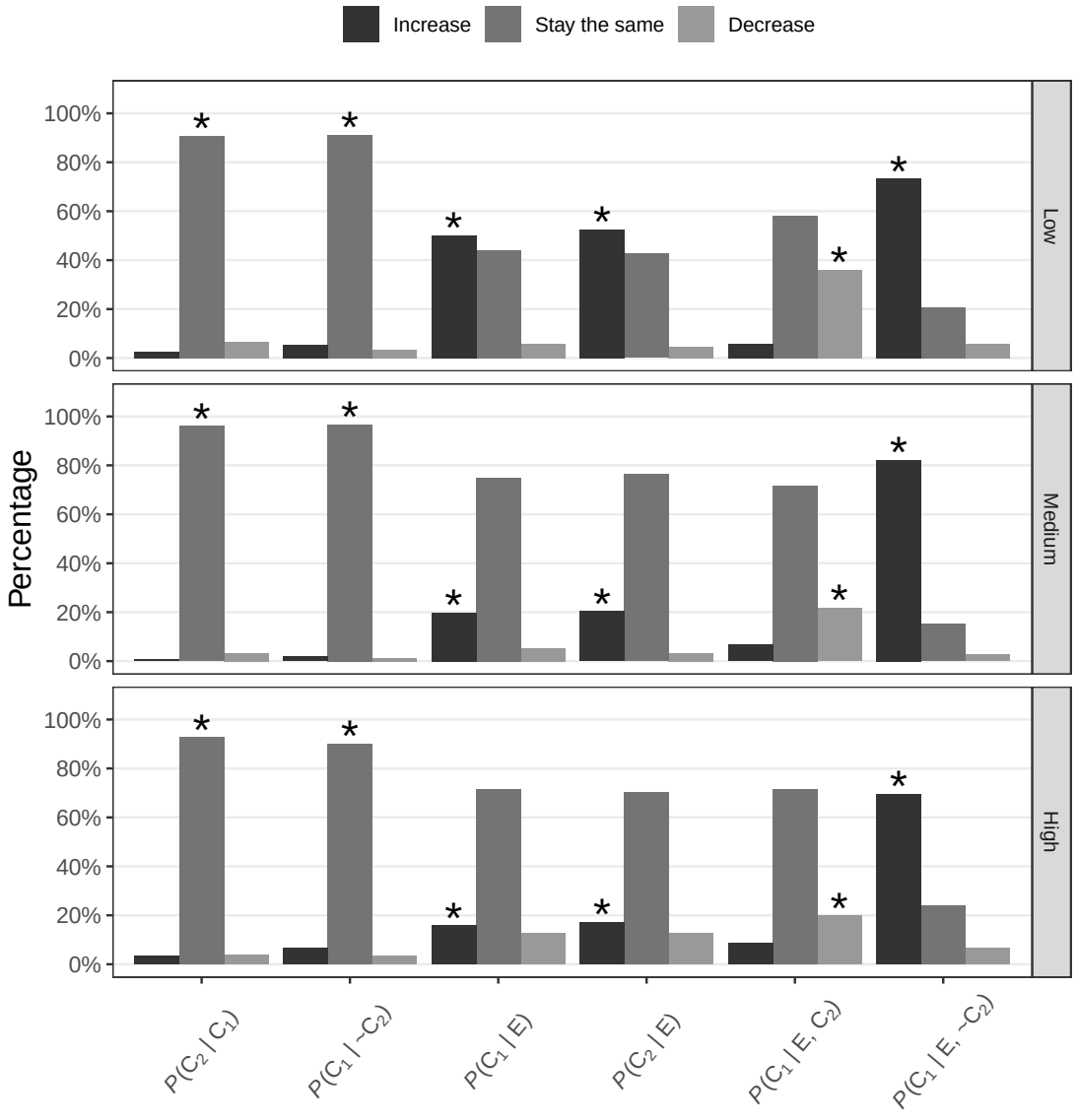


Figure 2.11: Distribution of participants' responses to qualitative questions in Experiment 1. Asterisks above the bars indicate normative answers.

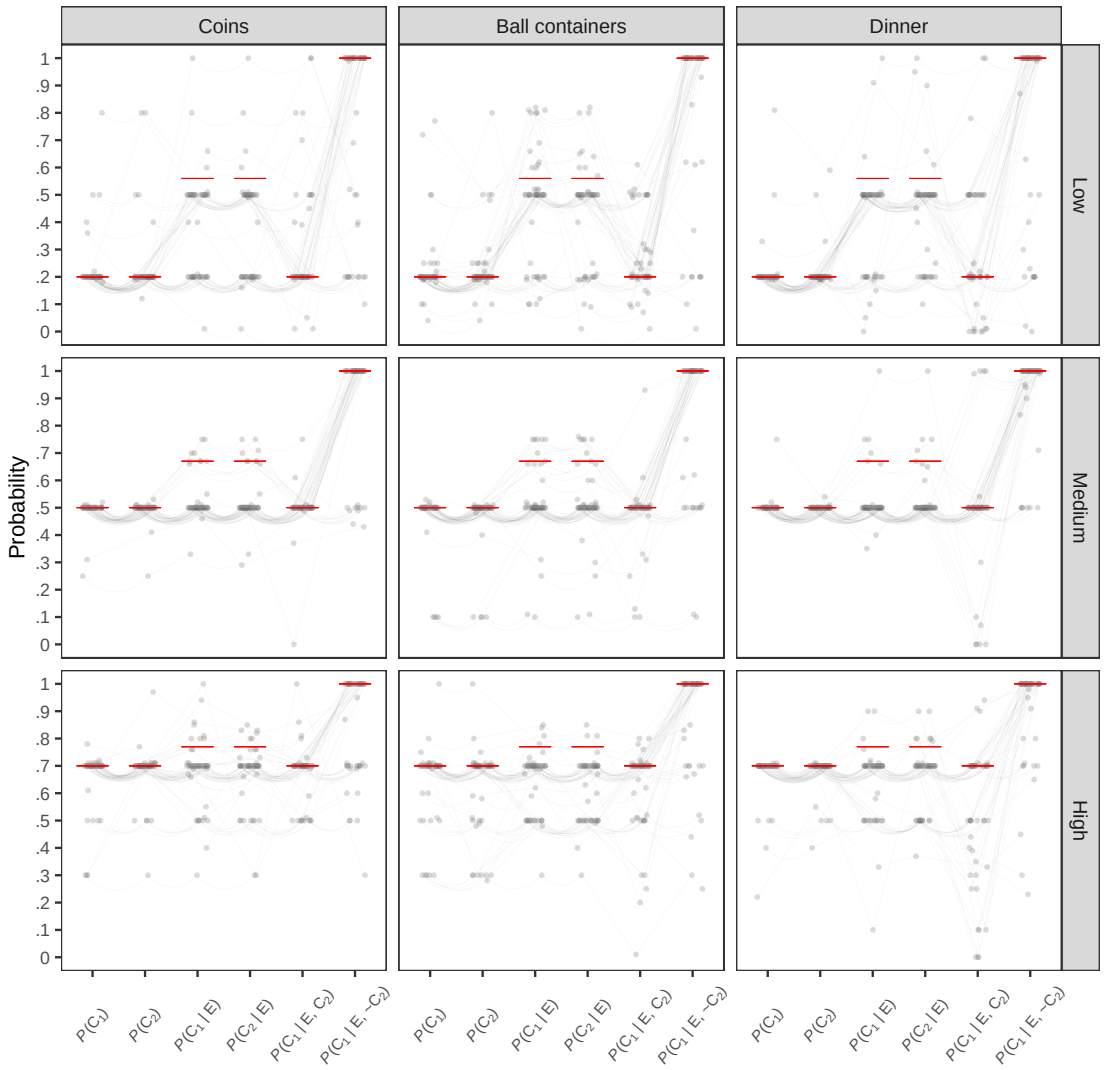


Figure 2.12: Participants' responses to quantitative questions in Experiment 1. Red horizontal lines are correct (normative) answers. Gray lines between data points depict how participants changed their probability estimates from one questions to another, with curved lines indicating that a participant did not change (within $\pm .02$) their probability estimate.

A correct quantitative estimate Before I proceed to the main analysis, I will first address what is meant by a correct quantitative estimate as this notion is important for the main analysis. All the quantitative estimates are provided on a scale from 0% to 100% using a slider with increments of 1%. Given that these estimates are entered using a slider, it is plausible that some variability in participants' quantitative estimates could be due to 'a slip of the hand'. This would happen if a participant wanted to enter a specific estimate (e.g. 20%), but instead entered different one that is in close proximity of the one they wanted to enter (e.g. 21%) because their hand slightly slipped when they were entering an estimate using a slider.

The variability that is due to a slip of the hand is, however, expected to be small and withing a range of ϵ that is taking a small value. Nonetheless, one would need to choose an appropriate value for ϵ in order to account for this variability.

To find the most appropriate ϵ I focused on participants' responses on the two priors questions. The priors questions are easiest to answer (in two cover stories correctly answering the two priors questions meant repeating the priors provided in the cover story) and any variability around the correct answer to the priors questions is most likely due to a slip of the hand. Thus, accounting for such variability would mean accounting for the slip of the hand.

To choose an ϵ, I have (i) selected of participants' responses that are within .05 of the correct answer as I expected the variation that is due to the slip of the hand to be small, specifically less than .05; (ii) of these selected responses I have subtracted the correct prior so that all the responses are comparable across the different priors; and (iii) I have calculated the variance of this data for each

of $0 \leq \epsilon \leq .05$ where, for example, $\epsilon = .03$ meant that all responses that are within $\pm.03$ of the correct prior are considered as the correct responses, i.e. the difference between these responses and the correct responses was assumed to be 0.

From Figure 2.13 we can see that the variance is quite small (the highest variance is around .5) and that there is a significant drop variance from $\epsilon = 0$ to $\epsilon = .02$ that then flattens out between $\epsilon = .02$ and $\epsilon = .04$, with again a significant drop between $\epsilon = .04$ and $\epsilon = .05$. This suggested that estimates that are equal to $\pm.05$ of the correct answer are plausibly not due to the slip of the hand; rather participants seemed to have aimed to provide an estimate that is different than the correct one, namely .55. Further, the graph also suggested that an appropriate ϵ should be between .02 and .04 as that is where the variance flattens out.

I have chosen ϵ to be equal to .02 in the further analyses for the following reasons. First, from Figure 2.13 there does not seem to be much different in variance for $.02 \leq \epsilon \leq .04$, so $\epsilon = .02$ already captures similar amounts of variance that is dues to the slip of the hand as $\epsilon = .03$ and $\epsilon = .04$. Second, a normative response in diagnostic reasoning for the high priors conditions is .77. If a participant mistakenly provided .73 as their estimate but they actually wanted to provide .7, this would have counted as a correct answer to the diagnostic reasoning question with $\epsilon = .04$. Similar situation arises when $\epsilon = .03$ and a participants provided .53 (with an intention of providing .5) as their estimate in diagnostic reasoning with low priors where the normative answer is .56. Third, only $\epsilon = .02$ is able to always distinguish between (i) estimates that are in line with the normative answers in both diagnostic reasoning in high

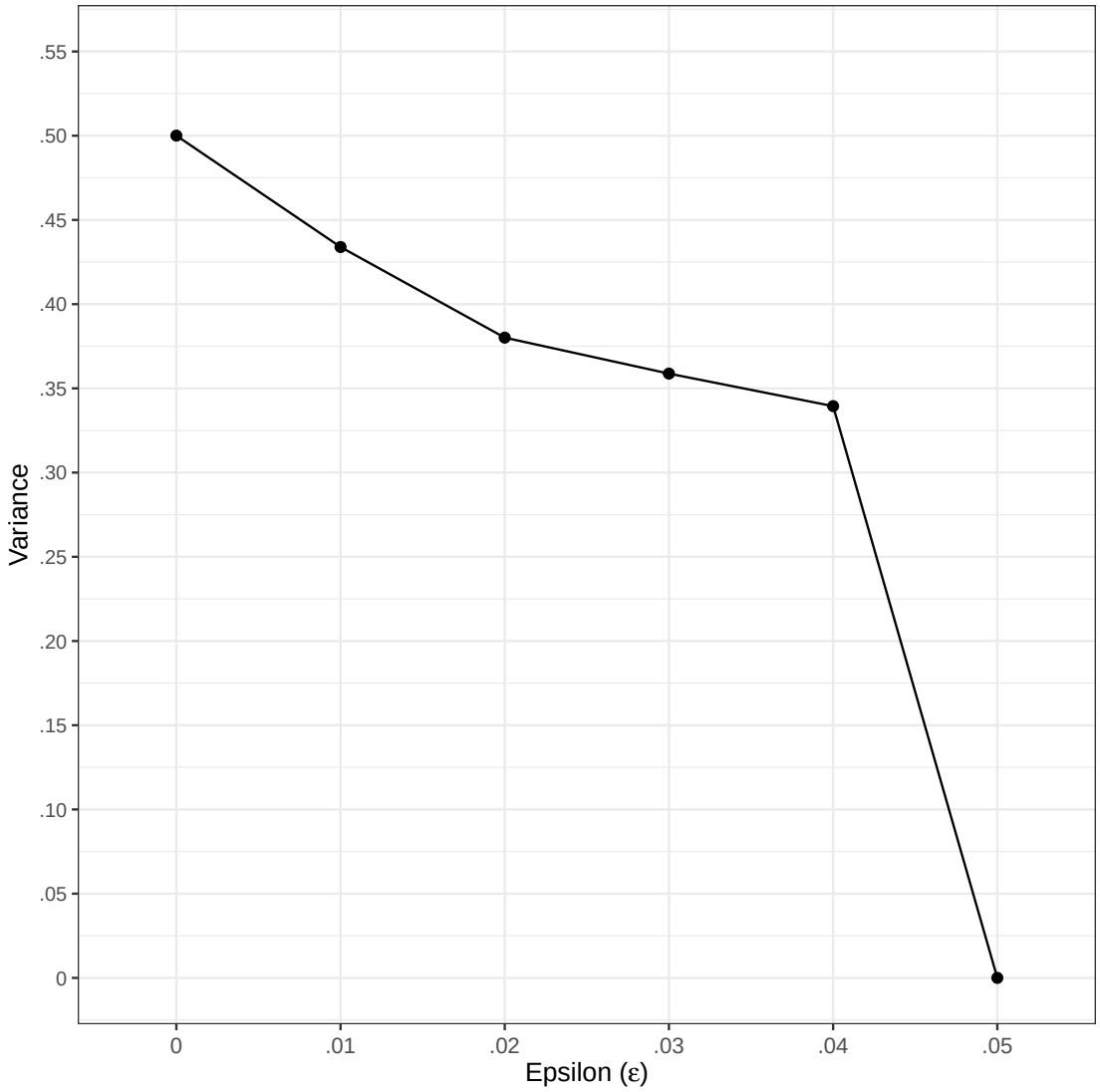


Figure 2.13: Variance in participants responses to the priors questions who provided estimates that are within $\pm.05$ of the correct answer as a function of different epsilons (ϵ).

(.77) and low (.56) priors conditions and (ii) estimates that are in line with the prediction of the propensity hypothesis in the high priors conditions (.7) and the prediction of the diagnostic split hypothesis in the low priors conditions (.5), without some estimates falling in the both (i) and (ii).

After choosing an appropriate value for ϵ, we can now move on to the main analyses.

Overall performance To test for a main effect of cover story and/or priors on participants' judgment accuracy I initially coded all participants' answers as correct (1) or incorrect (0). For all quantitative estimates, an answer was considered correct if it fell within $\pm$ .02 of the normative probability estimate. This allowed for a comparative measure of participants' accuracy for both qualitative and quantitative types of inferences. Subsequently, if an inference 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 2.1) I 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 leaves eight coded question-types regarding: priors, independence, qualitative diagnostic reasoning, quantitative diagnostic reasoning, qualitative explaining away, quantitative explaining away, qualitative logic, and quantitative logic.

To test the effect of Cover story and Priors on participants' overall performance (in the coded form) on the eight question-types, I 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). I found a main effect of Priors, $z = -3$, $p = .003$ and no main effect of Cover story, $z = 0.56$, $p = .58$. I 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 I found no main effect of Cover story on accuracy nor an interaction between Cover story and Priors, I collapsed data across cover stories to perform the subsequent analyses regarding participants' performance on explaining away. Therefore, I now compared across three groups: a low priors group ($\text{Group}_{\text{LOW}}$, $N = 150$), a medium priors group ($\text{Group}_{\text{MEDIUM}}$, $N = 152$), and a high priors group ($\text{Group}_{\text{HIGH}}$, $N = 151$).

Prior probabilities 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 $q^* = 0.05$[21] indicated a significant difference between $\text{Group}_{\text{MEDIUM}}$ (92.8%) and $\text{Group}_{\text{HIGH}}$ (78.1%), corrected $p = .002$. No significant difference was found between $\text{Group}_{\text{LOW}}$ (84.7%) and either $\text{Group}_{\text{MED}}$, corrected $p = .062$, and $\text{Group}_{\text{HIGH}}$, corrected $p = .192$.

Overall, 85.2% of participants across all conditions answered the priors correctly. I have also computed a sample standard deviation from the correct priors (s_{priors}) for each group. On $P(\text{C}_1)$ question, for $\text{Group}_{\text{LOW}}$, $s_{priors} = .11$,

[21] The same applied to all other pairwise comparisons in this part of the chapter.

95% CI [.08, .16][22]; for $\text{Group}_{\text{MEDIUM}}$, $s_{priors} = .07$, 95% CI [.05, .11]; for $\text{Group}_{\text{HIGH}}$, $s_{priors} = .11$, 95% CI [.1, .15]. On $P(\text{C}_2)$ question, for $\text{Group}_{\text{LOW}}$, $s_{priors} = .11$, 95% CI [.08, .16]; for $\text{Group}_{\text{MEDIUM}}$, $s_{priors} = .07$, 95% CI [.04, .11]; for $\text{Group}_{\text{HIGH}}$, $s_{priors} = .11$, 95% CI [.1, .15]. These results indicate relatively low standard deviations from the stated priors.

Therefore, although the difference between the above groups was significant, the high proportion of participants who stated the correct priors for both causes and the low deviation 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 $P(\text{C}_1)$ and $P(\text{C}_2)$ in Figure 2.12).

Independence of causes For a breakdown of the frequency of participants' choices on independence questions see Figure 2.11. Within each group I obtained the percentage of people who correctly answered *both* questions regarding the independence of causes (Q3 and Q4 in Table 2.1). Within $\text{Group}_{\text{LOW}}$ this was 88.7%, within $\text{Group}_{\text{MEDIUM}}$ this was 95.4% and within $\text{Group}_{\text{HIGH}}$ 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.

Diagnostic reasoning Independent analyses were conducted on qualitative and quantitative diagnostic reasoning questions (Qs 5–8 in Table 2.1).

[22] As the data are quite clearly non-normally distributed, 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).

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 2.11 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, 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 normative quantitative differences 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 2.12).

To gauge how much participants deviated from the normative estimates, I computed a sample standard deviation from the normative response (s_{norm}) for each group. On $P(\text{C}_1 \mid \text{E})$ question, for $\text{Group}_{\text{LOW}}$, $s_{norm} = .25$, 95% CI [.22, .27]; for $\text{Group}_{\text{MEDIUM}}$, $s_{norm} = .18$, 95% CI [.17, .2]; for $\text{Group}_{\text{HIGH}}$, $s_{norm} = .17$, 95% CI [.15, .2]. On $P(\text{C}_2 \mid \text{E})$ question, for $\text{Group}_{\text{LOW}}$, $s_{norm} = .24$, 95% CI [.22, .27]; for $\text{Group}_{\text{MEDIUM}}$, $s_{norm} = .18$, 95% CI [.16, .20]; for $\text{Group}_{\text{HIGH}}$, $s_{norm} = .17$, 95% CI [.15, .19]. This suggests that $\text{Group}_{\text{LOW}}$ most deviated from the normative answers compared to the other two groups.

I 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 I 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 2.2 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 estimates since the null hypothesis that the normative difference is

Table 2.2: Quantitative differences in diagnostic reasoning inferences per group in Experiment 1.

Inferences	Normative diff.	Empirical diff. 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	$< .001$

equal to the empirical difference is strongly rejected in all cases. Furthermore, only in $\text{Group}_{\text{LOW}}$ did the empirical difference go in the normative direction. In both $\text{Group}_{\text{MEDIUM}}$ and $\text{Group}_{\text{HIGH}}$ 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.

Direct explaining away Independent analyses were conducted on qualitative and quantitative questions regarding direct explaining away (Q9 and Q10 in Table 2.1).

Qualitative For a breakdown of the frequency of participants' choices on the qualitative direct explaining away question see Figure 2.11. 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 2.1.7.3), post-hoc pairwise comparisons using the FDR procedure illustrated a significant difference between $\text{Group}_{\text{LOW}}$ (36%) and both $\text{Group}_{\text{MEDIUM}}$ (21.7%), corrected $p = .013$ and $\text{Group}_{\text{HIGH}}$ (19.9%), corrected $p = .008$. No significant difference was found between $\text{Group}_{\text{MEDIUM}}$ and $\text{Group}_{\text{HIGH}}$, corrected $p = .8$. This suggests that participants in $\text{Group}_{\text{LOW}}$ performed significantly better than participants in $\text{Group}_{\text{MEDIUM}}$ and participants in $\text{Group}_{\text{HIGH}}$. 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 $\text{Group}_{\text{MEDIUM}}$ (82.9%), and both $\text{Group}_{\text{LOW}}$ (52.7%), corrected $p < .001$ and $\text{Group}_{\text{HIGH}}$ (57.6%), corrected $p < .001$. No significant difference was found between $\text{Group}_{\text{LOW}}$ and $\text{Group}_{\text{HIGH}}$, 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 I also computed a sample standard deviation from the normative response (s_{norm}). For $\text{Group}_{\text{LOW}}$, $s_{norm} = .22$, 95% CI [.19, .27]; for $\text{Group}_{\text{MEDIUM}}$, $s_{norm} = .15$, 95% CI [.12, .19]; for $\text{Group}_{\text{HIGH}}$, $s_{norm} = .21$, 95% CI [.17, .25]. This suggests that $\text{Group}_{\text{MEDIUM}}$ 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 the finding of overall poor performance on qualitative direct explaining away, a quick look at Figure 2.12 reveals that a large number of participants repeated the priors in $P(\text{C}_1 \mid \text{E})$, $P(\text{C}_2 \mid \text{E})$, and $P(\text{C}_1 \mid \text{E}, \text{C}_2)$ (this is discussed in Section 2.1.7.3 below). Since in the study $P(\text{C}_1) = P(\text{C}_1 \mid \text{E}, \text{C}_2)$ and a large proportion of participants did accept the priors (see Section 2.1.7.3), 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.

Logic Independent analyses were conducted on qualitative and quantitative 'logic' questions (Q11 and Q12 in Table 2.1).

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 $\text{Group}_{\text{MEDIUM}}$ (82.2%) and $\text{Group}_{\text{HIGH}}$ (69.5%), corrected $p = .043$. No significant difference was found between $\text{Group}_{\text{LOW}}$

(73.3%) and both $\text{Group}_{\text{MEDIUM}}$, corrected $p = .127$, and $\text{Group}_{\text{HIGH}}$, corrected $p = .548$. As can be seen from Figure 2.11, 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: $\text{Group}_{\text{LOW}}$, 68.7%; $\text{Group}_{\text{MEDIUM}}$, 77%; and $\text{Group}_{\text{HIGH}}$, 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.

Explaining away: relational concept Given the relational nature of explaining away, to better investigate participants' updating behaviour across this pattern of inference, I 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, I conducted the analysis on questions relating to direct explaining away and logic (Q9 and Q11 in Table 2.1).[23] A Chi-Square test of independence il-

[23]I did not include the two qualitative diagnostic reasoning questions here since these two questions are about the relationship between the priors and diagnostic reasoning. The aim was

lustrated 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 $\text{Group}_{\text{LOW}}$ (32.7%) and $\text{Group}_{\text{HIGH}}$ (15.9%), corrected $p = .003$ and between $\text{Group}_{\text{LOW}}$ and $\text{Group}_{\text{MEDIUM}}$ (20.4%), corrected $p = .033$. No significant difference was found between $\text{Group}_{\text{MEDIUM}}$ and $\text{Group}_{\text{HIGH}}$, corrected $p = .386$. Similarly to the qualitative diagnostic 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 I included in the analyses were those relating to the updating of C_1, namely, $P(\text{C}_1 \mid \text{E})$, $P(\text{C}_1 \mid \text{E}, \text{C}_2)$, and $P(\text{C}_1 \mid \text{E}, \sim\text{C}_2)$. These are Q6, Q10 and Q12 in Table 2.1.

A Friedman's ANOVA was carried out on participants' estimates of the quantitative relational explaining away questions, within each of the groups (see Figure 2.14). Results illustrated a significant difference between these estimates within $\text{Group}_{\text{LOW}}$, $\chi^2(2) = 155.9$, $p < .001$, within $\text{Group}_{\text{MEDIUM}}$, $\chi^2(2) = 190.9$, $p < .001$ and within $\text{Group}_{\text{HIGH}}$, $\chi^2(2) = 157.2$, $p < .001$.

Wilcoxon signed-rank tests were carried out to compare participants' estimates with normative ones (see Table 2.3 below). In each of the tests, the null

to analyze participants understanding of the inequalities in (2.3) which are about the relations between diagnostic reasoning and direct explaining away (Q9) and between direct explaining away and 'logic' (Q11).

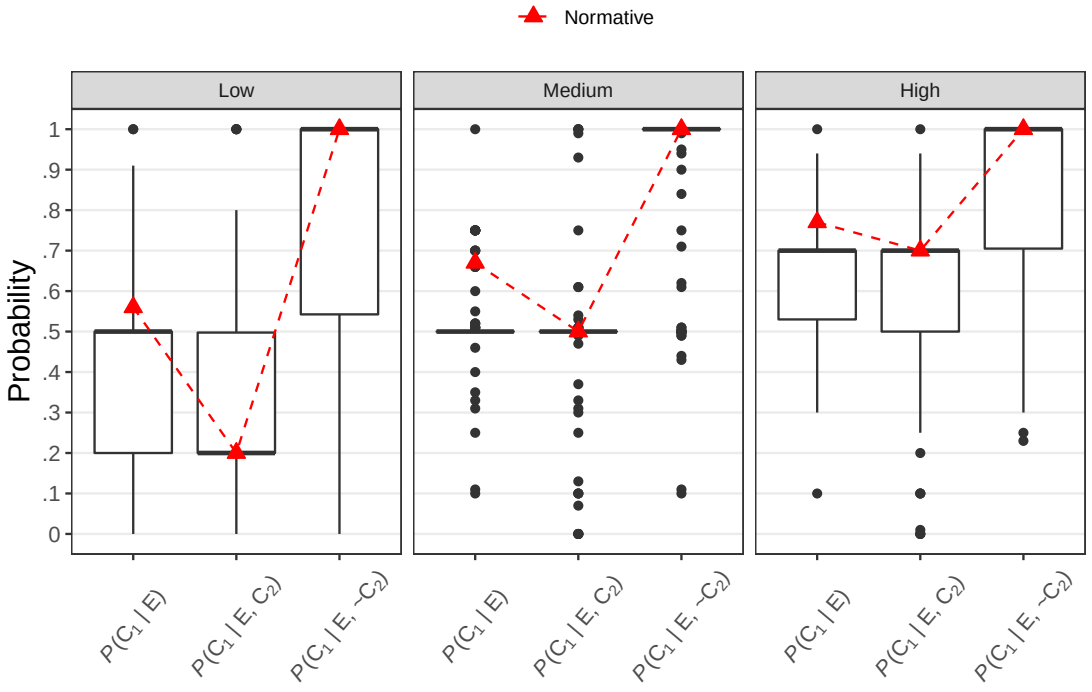


Figure 2.14: Box plots of participants' quantitative relational explaining away responses in three groups along with the normative estimates in Experiment 1.

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 $\text{Group}_{\text{HIGH}}$ between $P(\text{C}_1 \mid \text{E}, \text{C}_2)$ and $P(\text{C}_1 \mid \text{E}, \sim\text{C}_2)$ where participants appear to have sufficiently shifted their estimates. The participants in $\text{Group}_{\text{LOW}}$ and $\text{Group}_{\text{MEDIUM}}$ have thus under-adjusted their estimates despite the difference in medians between $P(\text{C}_1 \mid \text{E}, \text{C}_2)$ and $P(\text{C}_1 \mid \text{E}, \sim\text{C}_2)$ being equal to the normative difference for these groups.

Table 2.3: Within-group explaining away in Experiment 1.

Inferences	Normative diff.	Empirical diff. of medians	p-value
$Group_{\text{LOW}}$			
A − B	.36	.3	$< .001$
C − B	.8	.8	$< .001$
$Group_{\text{MED}}$			
A − B	.17	0	$< .001$
C − B	.5	.5	$< .001$
$Group_{\text{HIGH}}$			
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 := $P(\text{C}_1 \mid \text{E}, \sim\text{C}_2)$.

Diagnostic split To test the diagnostic split hypothesis I included in the analysis only participants who reported the correct priors ($N = 386$, or 85.2% of all participants) and then calculated the proportion of these participants who reported .5 ($\pm .02$) as their estimate for *both* $P(\text{C}_1 \mid \text{E})$ and $P(\text{C}_2 \mid \text{E})$. Of 386 participants, 50.4% in $\text{Group}_{\text{LOW}}$, 78.7% in $\text{Group}_{\text{MED}}$ and 13.6% in $\text{Group}_{\text{HIGH}}$ provided estimates in line with the diagnostic split hypothesis. 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 pro-

portions 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 $\text{Group}_{\text{MED}}$ is expected as the .5 response is predicted by both hypotheses. The relatively low proportion of participants observed in $\text{Group}_{\text{HIGH}}$ 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 this part of the chapter, I 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 I built simple cross-tabulations. I selected only participants who correctly answered the both priors questions ($N = 386$) and collapsed the data across all conditions. I 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 2.4) (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).[24] First, notice that the cross-tabulations for both diagnostic

[24] I have not included the diagnostic split hypothesis in cross-tabulations that included *qualitative* diagnostic reasoning and *qualitative* relational explaining away, as I did with the propensity hypothesis, since (i) the propensity hypothesis has a very specific quantitative prediction

reasoning and explaining away 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 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.

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

Qualitative I calculated the proportions of participants who, having stated the correct priors ($N = 386$), 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 (of $N = 386$): 63.8% for $\text{Group}_{\text{COINS}}$, 53.8% for $\text{Group}_{\text{BALL\_CONTAINERS}}$, and 46.8% for $\text{Group}_{\text{DINNER}}$. A Chi-Square test of in-

that does not depend 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 I have not included the diagnostic split hypothesis in cross-tabulations on the qualitative results.

Table 2.4: 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. Total $N = 386$.

		Diag. reasoning		Explaining away	
		Quantitative		Quant. relational	
		Yes	*No*	*Yes*	*No*
Diag. 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

dependence 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 $\text{Group}_{\text{COINS}}$ and $\text{Group}_{\text{DINNER}}$, corrected $p = .021$. No significant difference was found between $\text{Group}_{\text{COINS}}$ and $\text{Group}_{\text{BALL\_CONTAINERS}}$, corrected $p = .213$, or between $\text{Group}_{\text{BALL\_CONTAINERS}}$ and $\text{Group}_{\text{DINNER}}$, corrected $p = .316$.

Quantitative Out of the participants who correctly stated the priors, I calculated the proportions of those who provided the priors as their estimate

to $P(C_1 \mid E)$, $P(C_2 \mid E)$, and $P(C_1 \mid E, C_2)$ (i.e. Q6, Q8, and Q10). Collapsing across the priors conditions, the percentages were (of $N = 386$): 60.8% for $\text{Group}_{\text{COINS}}$, 50% for $\text{Group}_{\text{BALL\_CONTAINERS}}$ and 44.6% for $\text{Group}_{\text{DINNER}}$. 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 $\text{Group}_{\text{COINS}}$ and $\text{Group}_{\text{DINNER}}$, corrected $p = .034$. No significant difference was found between $\text{Group}_{\text{COINS}}$ and $\text{Group}_{\text{BALL\_CONTAINERS}}$, corrected $p = .198$, or between $\text{Group}_{\text{BALL\_CONTAINERS}}$ and $\text{Group}_{\text{DINNER}}$, corrected $p = .421$.

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

Furthermore, from Table 2.4 we can see that the propensity hypothesis accounted for about 53% of violations in both the quantitative diagnostic reasoning and quantitative relational explaining away (of those who correctly answered both priors questions). I 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 2.5 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 (of $N = 386$). The high percentages suggest that the

propensity hypothesis was driving the majority of violations in all these inferences. Table 2.6 further elucidates the point from Section 2.1.7.3 where I 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 2.5: 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. Total $N = 386$.

	Diag. 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

2.1.7.4 Discussion

The methodology I used in Experiment 1 has resulted in the large proportions of participants accepting the priors given to them, not violating the indepen-

Table 2.6: 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

dence 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 results, the 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 I 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. I have found that participants' quantitative responses only partially supported this prediction: only in diagnostic reasoning I have found that the difference $P(C_i \mid E) - P(C_i)$ 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, I 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 the two hypotheses.

As predicted by the propensity interpretation hypothesis, I found that

a significant proportion of participants reported that $P(C_i) = P(C_i \mid E) = P(C_i \mid E, C_j)$ in both qualitative and quantitative questions. Moreover, I found that this proportion was the highest when participants were reasoning with the cover story in which I expected the propensity interpretation to be the most pronounced (Coins cover story) and the lowest when participants reasoned with the cover story in which I 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) (of those who correctly answered the priors, $N = 386$).

Finally, regarding the diagnostic split hypothesis I 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 (of $N = 386$). 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 2.4 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 (of $N = 386$). Given this and the other above-mentioned high percentages, I 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. This result, however, also suggests that there is around 22% of the violations in quantitative diagnostic reasoning and relational explaining away (total $N = 386$) that the two hypotheses were not able to capture. This percentage us high enough to suggest that there are factors at work other than the two hypotheses that drive participants estimates in diagnostic reasoning and explaining away. Further work should endeavor to identify these factors.

Before moving on to the next experiment I would like to address potential carry-over effects that might arise due to the order of the questions being kept constant across all conditions in Experiment 1 (subsequent experiments). Many previous studies on explaining away have elicited probabilities estimates from participants on different types of judgments (priors, independence, diagnostic reasoning, explaining away etc.) in a random order (see for example Fernbach & Rehder, 2013; Rehder, 2014a; Rehder & Waldmann, 2017; Rottman & Hastie, 2016). All these studies have, however, found similar results in that

people's (quantitative) diagnostic reasoning and explaining away judgements were poor. Given that the main results are similar, the set order of questions asked in this experiment (and those that follow it) does not seemed to have yielded in different performance on diagnostic reasoning and explaining away questions. That being said, it is possible that a high percentage of participants who correctly answered the priors questions could partly be attributed to the fact that this was the first question they were presented with rather than the question the order of which was random (many of the studies who used the random order of questions reported low percentage of participants who provided the correct priors), which would be an instance of the primacy effects. Given the importance of priors in the determining the amount of explaining away, these potential primacy effects would, however, be desirable in the contexts of this and the following experiments are they would potentially increase the acceptance of the priors.

2.1.8 Experiment 2[25]

Experiment 1 suggested that the propensity interpretation of probability and the diagnostic split strategy are some of the factors that are significantly driving the findings regarding reasoning in explaining away. One of the limitation of Experiment 1 was that it explored explaining away in situations where the prior probabilities of causes were equal. The goal of this experiment was to examine the robustness of the two hypotheses by testing them in explaining situations where the priors of causes were not equal.

The results from Experiment 1 suggested that largest differences between

[25]This experiment was conducted together with Alice Liefgreen (Liefgreen & Tešić, in press).

the predictions of the normative model and the participants' probability estimates were in the low priors condition. From the perspective of the two hypotheses this is expected as the predictions of these hypotheses diverge more from the normative ones when the priors are low: the normative amount of explaining away is larger when the priors are lower increasing the discrepancy between the normative predictions and the predictions of the two hypotheses. This implies that the power of the experiment is higher if the priors are low rather than medium or high. Therefore, all the scenarios in this next experiment had priors of the causes that were low.

In addition to being low, the priors in this experiment were unequal: $P(C_1) = .2$ and $P(C_2) = .1$. The predictions of the propensity hypothesis do not change due to unequal priors: it still predicts that $P(C_i) = P(C_i \mid E) = P(C_i \mid E, C_j)$, where $i, j \in \{1, 2\}$. Therefore, this experiment sought to replicate the findings of Experiment 1 regarding the propensity hypothesis. The diagnostic split hypothesis predicts that $P(C_1 \mid E) + P(C_2 \mid E) = 1$, so was also seeking to replicate the findings of Experiment 1. However, given that the priors are now unequal, one would expect participants not just to split the probability space equally between the two cases, i.e. $P(C_1 \mid E) = P(C_2 \mid E) = .5$, but, as the findings from Liefgreen et al. (2018) suggested, some participants could also follow the $2 : 1$ ratio between the priors and split the probability space accordingly in diagnostic reasoning, i.e. they would estimate that $P(C_1 \mid E) = .67$ and $P(C_2 \mid E) = .33$. Thus, two clusters of participants estimates are expected in this experiment that accord with the diagnostic split predictions.

2.1.8.1 Overview

The same three cover stories from Experiment 1 were used to manipulate how much the propensity interpretation was emphasized. A deterministic set-up was again adopted, i.e. the presence of at least one cause entailed the presence of the effect: $P(\mathrm{E} \mid \mathrm{C}_1, \mathrm{C}_2) = P(\mathrm{E} \mid \mathrm{C}_i, {\sim}\mathrm{C}_j) = 1$; and the absence of both causes entailed absence of the effect: $P(\mathrm{E} \mid {\sim}\mathrm{C}_1, {\sim}\mathrm{C}_2) = 0$. In this experiment, however, the priors were low and unequal with $P(\mathrm{C}_1) = .2$ and $P(\mathrm{C}_2) = .1$.

As the only manipulation in this experiment was how much the propensity was pronounced across the three cover stories, only the propensity hypothesis was directly tested in this experiment. However, as mentioned above the diagnostic split hypothesis has clear predictions for this experiment and I again cross-tabulated the data to explore how much of the violation in diagnostic reasoning can be explained by the diagnostic split hypothesis.

2.1.8.2 Methods

Participants and Design A total of 271 participants ($N_{\mathrm{MALE}} = 11$, 4 participants identified as 'other', $M_{\mathrm{AGE}} = 32.2$ 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 10.8 minutes to complete. Eight participants were excluded as they did not pass the attention check, leaving a total of 263 participants in the analyses.

A between-participant design was employed and participants were randomly allocated to one of three groups which differed in the cover story they were required to reason with: $\mathrm{Group}_{\mathrm{COIN}}$ ($n = 87$), $\mathrm{Group}_{\mathrm{BALL}}$ ($n = 87$), and

$\text{Group}_{\text{DINNER}}$ ($n = 89$).

Materials The materials used in this experiment were exactly the same as in Experiment 1, with the exception that in this experiment the priors were unequal and only low priors were employed. For full materials visit Open Science Framework, https://osf.io/zm6ec/.

Procedure The procedure for this experiment was exactly the same as for Experiment one: participants were asked the same number of questions, the same types of questions, and in the same order as in Experiment 1.

2.1.8.3 Results

Participants' answers to all qualitative questions in the inference questionnaire are represented in Figure 2.15 and their answers to all quantitative questions are in Figure 2.16. As in Experiment 1, a separate analyses was carried out for each inference type.

Overall Performance To determine the effect of the manipulations on participants' overall performance throughout the task, like in Experiment 1 a GLM with binomial link function was built. The model had one fixed effect, namely Cover story, with a random intercept for each participant (there was no random slope for participant since Cover story varies between participants). There was no main effect of Cover story, $z = -1.43$, $p = .15$. Including the predictor Cover story in the model did not improve model fit ($\chi^2(1) = 2.04$, $p = .15$) compared to just having an intercept as a predictor. As the predictor was centered, this implied that the data grand mean fits the data no worse than the model which

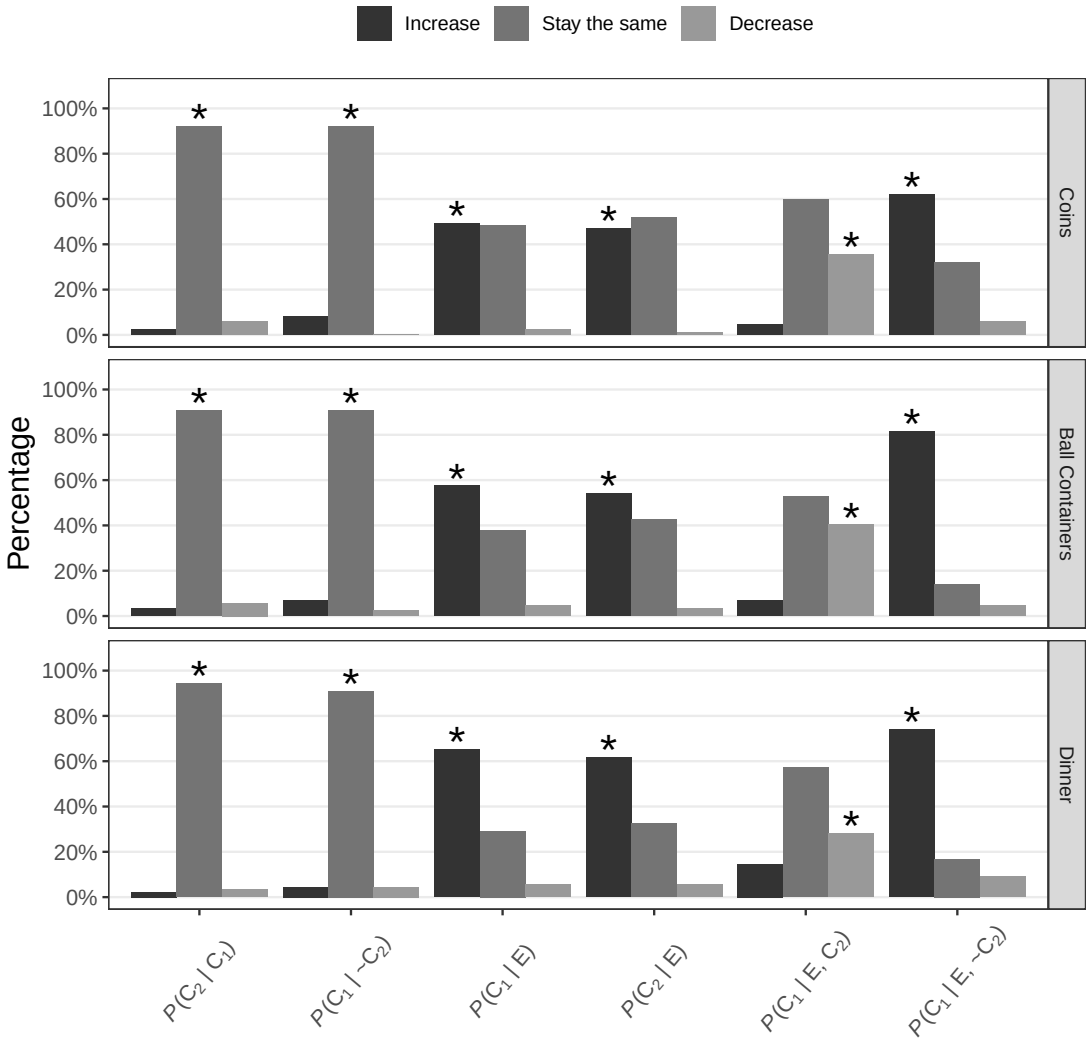


Figure 2.15: Distribution of participants' responses to qualitative questions in Experiment 2. Asterisks above the bars indicate normative answers.

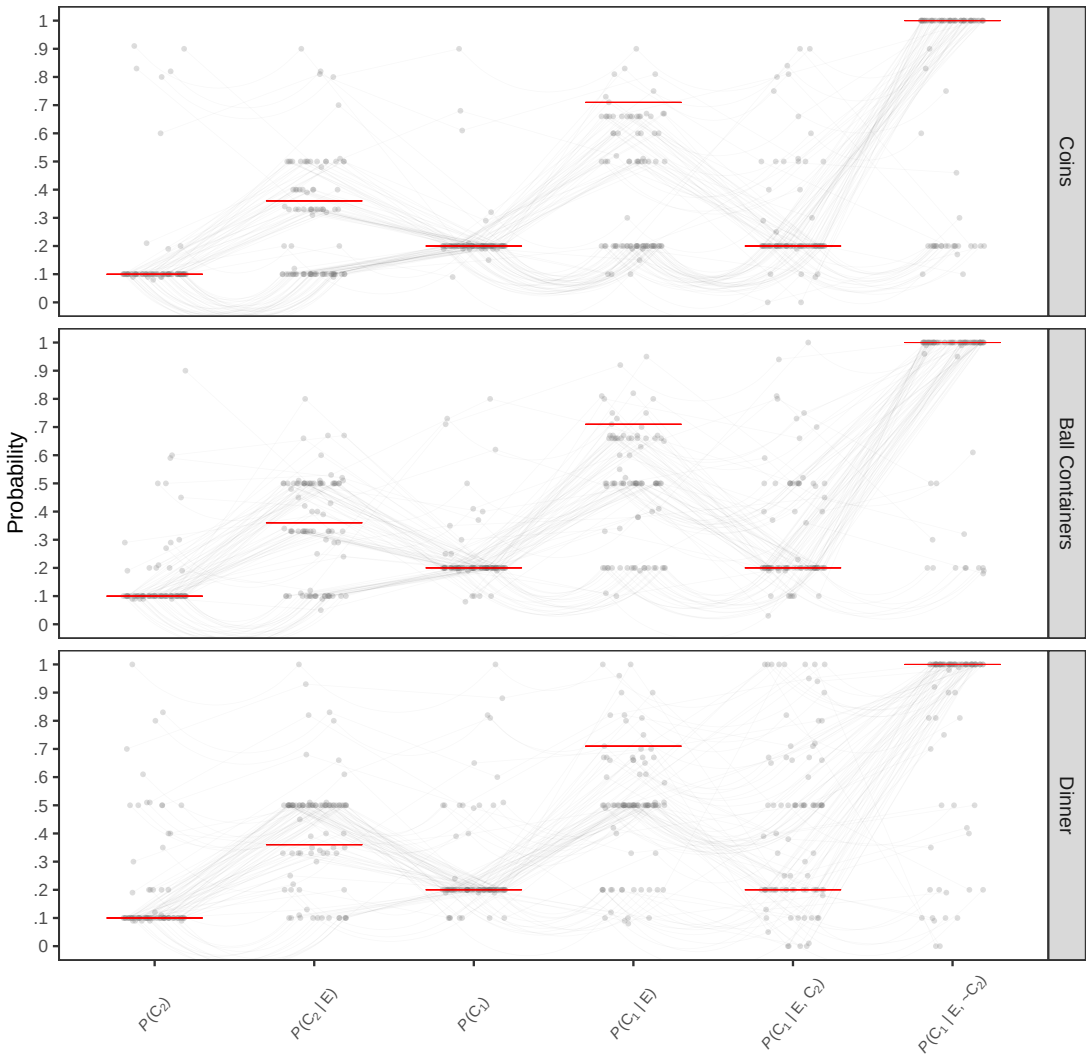


Figure 2.16: Participants' responses to quantitative questions in Experiment 2. Red horizontal lines are correct (normative) answers. Gray lines between data points depict how participants changed their probability estimates from one questions to another, with curved lines indicating that a participant did not change (within $\pm$ .02) their probability estimate. Note that the order of questions in this figure does not correspond to the order of question in the experiment. It starts with participants' estimates for $P(C_2)$ rather than $P(C_1)$. This is done purely to aid the visual inspection of the data as the priors of the two causes were not equal. Specifically, it aids the inspection of participants' estimates that did not change from one question to another.

includes the predictor.

Given there was no effect of Cover story, the data were collapsed across all three cover story conditions in order to obtain the following descriptives regarding participants' accuracy.

Prior probabilities Collapsing across all conditions, 79.1% of participants correctly answered *both* questions pertaining to the prior probabilities, i.e. $P(C_1)$ and $P(C_2)$. This is in line with the findings from Experiment 1 and it suggests that the majority of participants accepted the priors.

Independence of causes For a breakdown of the frequency of participants' answers to qualitative independence questions see Figure 2.15. Collapsing across conditions, 89% of participants correctly answered *both* questions relating to independence, i.e. Qs $P(C_2 \mid C_1)$ and $P(C_1 \mid \sim C_2)$. This is again very much in agreement with the findings from Experiment 1 and suggests that participants did not violate the assumption of Independence between the cases (before knowing the evidence).

Diagnostic Reasoning As in Experiment 1, both *qualitative* and *quantitative* estimates for diagnostic reasoning were collected. 50.2% of participants correctly answered *both qualitative* diagnostic reasoning questions and only .4% of participants (i.e. only 1 participant) correctly answered *both quantitative* diagnostic reasoning questions, i.e. $P(C_1 \mid E)$ and $P(C_2 \mid E)$. This again very closely mirrors the finding from Experiment 1 $\text{Group}_{\text{LOW}}$, i.e. the group that reasoned with low priors. The amount and direction of change in participants' quantitative estimates from their given priors to their estimates after learning about

the effect was also explored. As in Experiment 1, I conduced the Wilcoxon signed-rank test on the difference between participants' quantitative estimates on each prior question and their related diagnostic reasoning question (i.e. between $P(C_1)$ and $P(C_1 \mid E)$ and between $P(C_2)$ and $P(C_2 \mid E)$). Table 2.7 shows the normative differences, the empirical differences of medians, and p-values of Wilcoxon signed-rank tests.

Table 2.7: Quantitative differences in diagnostic reasoning inferences per group in Experiment 2.

Inferences	Normative diff.	Empirical diff. of medians	p-value
$P(C_1 \mid E) - P(C_1)$	.51	.3	$< .001$
$P(C_2 \mid E) - P(C_2)$	.26	.25	$< .001$

As in Experiment 1, the results here also suggest that participants have under-adjusted their probability estimates and that difference went in the normative direction. However, as the priors are unequal in this experiment, the normative difference between $P(C_1 \mid E)$ and $P(C_1)$ was higher than in Experiment 1, implying that from the normative perspective participants even more under-adjusted the probability $P(C_1 \mid E)$ than in Experiment 1. Analogously, since the normative difference between $P(C_2 \mid E)$ and $P(C_2)$ was lower than in Experiment 1, participants' under-adjustment of $P(C_2 \mid E)$ was lower than in Experiment 1. The specific clusters of participants' estimates in diagnostic reasoning are discussed in more detail below.

Direct explaining away 34.6% of participants correctly answered the *qualitative* direct explaining question (i.e. $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$) and 51% participants correctly answered the *quantitative* direct explaining question. I have also computed a sample standard deviation from the normative response (s_{norm}) using the same bootstrapping procedure as in Experiment 1; $s_{norm} = .28$, 95% CI [.25, .32]. These results are again on par with those in Experiment 1, Group_{LOW}. Again, these higher percentages again suggest a relatively better performance on direct quantitative explaining away. However, a quick look at Figure 2.16 reveals that a large number of participants repeated the priors in $P(\mathrm{C}_1 \mid \mathrm{E})$, $P(\mathrm{C}_2 \mid \mathrm{E})$, and $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ and thus 'correctly' answered the direct explaining question. This is further discussed below.

Logic 72.6% of participants correctly answered the *qualitative* direct explaining question (i.e. $P(\mathrm{C}_1 \mid \mathrm{E}, \sim\mathrm{C}_2)$) and 73.3% participants correctly answered the *quantitative* direct explaining question. These results are again on pair with those in Experiment 1 and suggest that participant largely understood the (deterministic) relationships in the cover stories.

Explaining away: relational concept As in the analysis in Experiment 1, participants' understanding of explaining away as a relational concepts was explored. 31.6% participants correctly answered *both* questions related to the *qualitative* relational explaining away, i.e. $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ and $P(\mathrm{C}_1 \mid \mathrm{E}, \sim\mathrm{C}_2)$.

A Friedman's ANOVA was carried out on participants' estimates of the *quantitative* relational explaining away questions: i.e. $P(\mathrm{C}_1 \mid \mathrm{E})$, $P(\mathrm{C}_1 \mid \mathrm{E}, \mathrm{C}_2)$ and $P(\mathrm{C}_1 \mid \mathrm{E}, \sim\mathrm{C}_2)$. Results illustrated a significant difference between these estimates: $\chi^2(2) = 265.6$, $p < .001$. Wilcoxon signed-rank tests were carried out

to compare participants' estimates with normative ones (see Table 2.8 below).

Table 2.8: Relational explaining away in Experiment 2.

Inferences	Normative diff.	Empirical diff. of medians	p-value
A − B	.51	.3	$< .001$
C − B	.8	.8	$< .001$

Note: A := $P(C_1 \mid E)$, B := $P(C_1 \mid E, C_2)$, C := $P(C_1 \mid E, \sim C_2)$.

These results are again very similar to those in Experiment 1, $\text{Group}_{\text{LOW}}$. They suggest that participants' understanding of explaining away as a relational concept was relatively poor compared to the normative answers and that they mostly under-adjusted their probability estimates.

Diagnostic split The predictions of the diagnostic split hypothesis was that participants would report their estimates in diagnostic reasoning such that $P(C_1 \mid E) + P(C_2 \mid E) = 1$. As in this experiment priors were unequal, it was expected that instead of just splitting the probability space equally between the two causes in diagnostic reasoning, a proportion of participants would report that $P(C_1 \mid E) = .67$ and $P(C_1 \mid E) = .33$ to reflect the $2:1$ ratio of the priors. To explore how much of the data can be explained by the diagnostic split hypothesis, I only included participants who reported correct priors ($N = 208$, or 79.1% of all data) and calculated the proportion of participants who reported .5 ($\pm .02$) as their estimate for *both* diagnostic reasoning questions or .67 ($\pm .02$) for $P(C_1 \mid E)$ and .33 ($\pm .02$) for $P(C_2 \mid E)$. This proportion was 47.1%, with 60.2% of these 47.1% reporting .5 as their estimates for both diagnostic reason-

ing question and the other 39.8% reporting .67 for $P(\mathrm{C}_1 \mid \mathrm{E})$ and .33 ($\pm$ .02) for $P(\mathrm{C}_2 \mid \mathrm{E})$. The results suggest that (i) a large proportion of participants' diagnostic reasoning estimates can be accounted for by the diagnostic split hypothesis (in line with Experiment 1, $\mathrm{Group}_{\mathrm{LOW}}$ where this proportion was 50.4%) and (ii) that some participants did follow the priors ratio and provided diagnostic reasoning estimates in line with that ratio.

As in Experiment 1, I have also cross-tabulated participants' responses to explore how much of the deviations from the normative model in (quantitative) diagnostic reasoning and (quantitative) relational explaining away can be accounted for by the diagnostic split hypothesis. From Table 2.9 we can see that about 47.1% of violation in quantitative diagnostic reasoning and 47.8% of violation in quantitative relational explaining away can be accounted for by the diagnostic split hypothesis (of $N = 208$).

Propensity interpretation To test the propensity hypothesis, I calculated the proportion of participants who did not update their probabilities when presented with evidence.

The proportions of participants who selected 'stay the same' as an answer to both *qualitative* diagnostic reasoning questions as well as the *qualitative* direct explaining away question for the three cover stories were ($N = 208$): 45.5% for $\mathrm{Group}_{\mathrm{COIN}}$, 34.3% for $\mathrm{Group}_{\mathrm{BALL}}$, and 21.9% for $\mathrm{Group}_{\mathrm{DINNER}}$. A Chi-Square test of independence showed a significant difference between these proportions, $\chi^2(2) = 8.59$, $p = .014$. Post-hoc pairwise comparisons using the FDR procedure showed that the only significant difference was between $\mathrm{Group}_{\mathrm{COIN}}$ and $\mathrm{Group}_{\mathrm{DINNER}}$, corrected $p = .018$. No significant difference was found between $\mathrm{Group}_{\mathrm{COIN}}$ and $\mathrm{Group}_{\mathrm{BALL}}$, corrected $p = .235$, or between $\mathrm{Group}_{\mathrm{BALL}}$

Table 2.9: 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 2. Total $N = 208$.

		Diag. reasoning		Explaining away	
		Quantitative		Quant. relational	
		Yes	*No*	*Yes*	*No*
Diag. split	Propensity interpretation (quantitative)				
Yes	*Yes*	0	0	0	0
Yes	*No*	0	98	0	98
No	*Yes*	0	57	0	57
No	*No*	0	53	3	50

and $\text{Group}_{\text{DINNER}}$, corrected $p = .235$.

To test the propensity hypothesis on participants' answers to *quantitative* questions, I calculated the proportions of participants who correctly stated the priors and who provided the priors as their estimate to $P(\text{C}_1 \mid \text{E})$, $P(\text{C}_2 \mid \text{E})$, and $P(\text{C}_1 \mid \text{E}, \text{C}_2)$ for the three cover stories (of $N = 208$): 40.3% for $\text{Group}_{\text{COIN}}$, 26.9% for $\text{Group}_{\text{BALL}}$, and 12.5% for $\text{Group}_{\text{DINNER}}$. A Chi-Square test of independence showed a significant difference between these proportions, $\chi^2(2) = 13.55$, $p = .001$. Post-hoc pairwise comparisons using the FDR

procedure showed that the only significant difference was between $\text{Group}_{\text{COIN}}$ and $\text{Group}_{\text{DINNER}}$, corrected $p = .002$. No significant difference was found between $\text{Group}_{\text{COIN}}$ and $\text{Group}_{\text{BALL}}$, corrected $p = .13$, or between $\text{Group}_{\text{BALL}}$ and $\text{Group}_{\text{DINNER}}$, corrected $p = .098$.

These results almost exactly replicate those from Experiment 1. Significantly more participants stayed at the priors in the Coins cover story where the propensity hypothesis was expected to be the most pronounced compared to the Dinner cover story, with the Ball containers cover story falling in between.

Furthermore, from Table 2.9 we can see that the propensity hypothesis accounted for about 27% of violations in both the quantitative diagnostic reasoning and quantitative relational explaining away (of $N = 208$). I 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 2.10 shows that the propensity hypothesis accounted for about 77% of violations in qualitative diagnostic reasoning, about 56% of violations in qualitative direct explaining away, and about 54% of violations in qualitative relational explaining away ($N = 208$). Similarly, to the findings in Experiment 1, these percentages suggest that the propensity hypothesis was driving the majority of violations in all these inferences. Table 2.11 further shows that about 44% of 'correct' responses in quantitative direct explaining away 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 2.10: 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 2. Total $N = 208$.

	Diag. reasoning		Qualitative explaining away			
	Qualitative		Direct		Relational	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*
Propensity interpretation (qualitative)						
Yes	0	72	0	72	0	72
No	114	22	79	57	75	61

2.1.8.4 Discussion

The results from Experiment 2 very closely resemble those from Experiment 1, in particular those for $\text{Group}_{\text{LOW}}$: relatively poor diagnostic reasoning and explaining away, with a large percent of participant under-adjusting their estimates compared to the normative ones. The findings, however, suggest that the deviations from the normative model can be accounted for by two hypotheses. A significant proportion of participants reported $P(\text{C}_\text{i}) = P(\text{C}_\text{i} \mid \text{E}) = P(\text{C}_\text{i} \mid \text{E}, \text{C}_\text{j})$ in both qualitative and quantitative questions, with the proportion of participant reporting these estimates being the highest in the Coins cover story where it was expected that the propensity interpretation would

Table 2.11: 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 2. Total $N = 208$.

	Quantitative direct explaining away	
	Yes	*No*
Propensity interpretation (quantitative)		
Yes	57	0
No	74	77

be the most pronounced and the lowest in the Dinner party cover story where the propensity interpretation was expected to be the least pronounced, with the Ball and containers cover story proportion falling in between these two. As in Experiment 1, these results support the hypothesis that people do interpret probabilities as propensities and that this can lead to deviations from the normative model in explaining away situations. The cross-tabulations further showed that the propensity hypotheses accounted for a significant percentage of violations in diagnostic reasoning and direct and relational explaining away (both qualitative and quantitative).

The diagnostic split hypothesis was also able to account for a large proportion of deviations from the normative account in quantitative diagnostic reasoning. In contrast to Experiment 1, in this experiment I found that some participants provided estimates that reflected the ratios of the priors of the

two causes. This was, however, expected as it was also suggested by the results from Liefgreen et al. (2018) and it fits the diagnostic split prediction that $P(C_1 \mid E) + P(C_2 \mid E) = 1$. One could argue that the diagnostic split predictions are under-specified as it allows $P(C_1 \mid E)$ and $P(C_2 \mid E)$ to assume any probability values as long as they add up to one. Two points here. First, the diagnostic split prediction that $P(C_1 \mid E)$ and $P(C_2 \mid E)$ have to add up to one to support the hypothesis is already a very specific one: any combination of $P(C_1 \mid E)$ and $P(C_2 \mid E)$ that does not add up to one would falsify the hypothesis. Second, the participants who seemed to have engaged in the diagnostic split reasoning did not provide estimates for the probability of the two causes after learning the effect that are random. In Experiment 1 they were mostly around .5 probability and Experiment 2 these were mostly around again .5 as well as .67 (for C_1) and .33 (for C_2). Is seems then that there are at least two factors that drive the estimates of participants who engage in diagnostic split reasoning: the number of causes and the ratio of the priors of the two causes. It remains to be seen what other factors contribute to how people choose estimates when engaging in diagnostic split reasoning. Nevertheless, the results from Experiment 1 and 2 suggest that they do not choose them at random.

Overall, the two hypotheses have accounted for around 75% of violations in quantitative diagnostic reasoning and relational explaining away (of those who answered both priors questions correctly, $N = 208$), thus accounting for a significant amount of the observed insufficiency in explaining away and very closely replicating the results from Experiment 1.

2.1.9 Experiment 3[26]

In Experiments 1 and 2, a significant proportion of participants provided .5 probability as their estimate 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 Experiments 1 and 2 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 3 was to disentangle the two possibilities and further extend results of Experiments 1 and 2 to more than 2 causes. To do so, in Experiment 3 I prompted participants to reason with a 4-node common-effect CBN with three causes (see Figure 2.17).

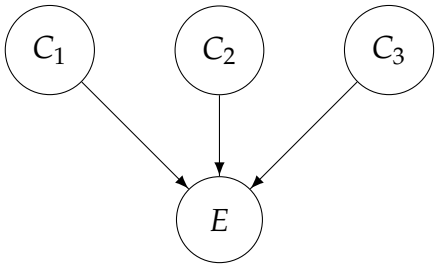


Figure 2.17: A common-effect CBN model with three causes.

In the CBN from Figure 2.17, assuming the deterministic set-up like in Experiments 1 and 2, the diagnostic split hypothesis would predict that $P(\mathrm{C}_1 \mid \mathrm{E}) +$

[26]This experiment was conducted together with Alice Liefgreen and David Lagnado (Tešić et al., 2020).

$P(C_2 \mid E) + P(C_3 \mid E) = 1$. If we further assume equal priors for all 3 causes, then if people engage in the diagnostic split reasoning we would expect that they provide the following estimates: $P(C_1 \mid E) = P(C_2 \mid E) = P(C_3 \mid E) = \frac{1}{3} \approx .33$. 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 $P(C_i \mid E)$ being lower than $P(C_i)$ (as was the case in High condition in Experiment 1), which is opposite to the normative direction of the update where $P(C_i \mid E) > P(C_i)$. In Experiment 1 I 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 3 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.

2.1.9.1 Overview

Similarly to Experiments 1, I manipulated the priors of causes and presented participants with different cover stories. I again employed a deterministic set-up where the presence of at least one cause entailed the presence of the ef-

fect: $P(E \mid C_1, C_2, C_3) = P(E \mid C_i, C_j, \sim C_k) = P(E \mid C_i, \sim C_j, \sim C_k) = 1$; and absence of all three causes entailed absence of the effect: $P(E \mid \sim C_1, \sim C_2, \sim C_3) = 0$. In this experiment, however, the priors were either low, $P(C_1) = P(C_2) = P(C_3) = .2$ or medium, $P(C_1) = P(C_2) = P(C_3) = .5$. I 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 $P(C_i \mid E) = .33 < .5 = P(C_i)$.

In this experiment I employed two cover stories from Experiments 1 and 2, one involving balls and containers, and one involving a dinner party. I did not use the cover story involving coin tossing since Experiments 1 and 2 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 3 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 3 I did not include the cover story including coin tossing.

Further, since in Experiments 1 and 2 the tests regarding the propensity hypothesis did not show significant difference between the balls and containers cover story and the dinner party cover story (although the ordinal difference was in line with the propensity hypothesis) I have not directly tested the propensity hypothesis in Experiment 3. However, given that the propensity hypothesis has a clear prediction in Experiment 3, namely $P(C_i \mid E) = P(C_i)$ for $i = \{1, 2, 3\}$, I again cross-tabulated the data to explore how much of the vi-

olation in diagnostic reasoning can be explained by the propensity hypothesis.

Given the new structure in Figure 2.17, 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').

2.1.9.2 Methods

Participants and Design A total of 119 participants ($N_{\text{MALE}} = 39$, 2 participants identified as 'other', $M_{\text{AGE}} = 35$ 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 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) $\times$ 2 (priors condition: low, medium) = 4 groups $N_{\text{BALL\_CONTAINERS\_LOW}} = 28$, $N_{\text{BALL\_CONTAINERS\_MED}} = 30$, $N_{\text{DINNER\_LOW}} = 32$, $N_{\text{DINNER\_MED}} = 29$).

Materials Each of the groups was asked to complete an inference questionnaire ($N_{\text{QUESTIONS}} = 12$), comprising of questions regarding priors and (un-

conditional) 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 2.12. 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 reasoned either with low or medium priors and were either presented the balls and containers cover story or the dinner party cover story from Experiments 1 and 2 now adapted to include the third cause (see Figure 2.18). For full materials visit Open Science Framework, https://osf.io/aqjkp/.

Procedure Like in Experiments 1 and 2, participants in each of the four groups were initially presented with the pertinent cover story and were given explicit information on the common-effect model embedded within the cover story including the prior probability of each cause, and the causal relationships within the model. This was done in both textual form and in visual form (graphical representation; see Figure 2.18). 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 2.12). 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 2.12 required participants to pro-

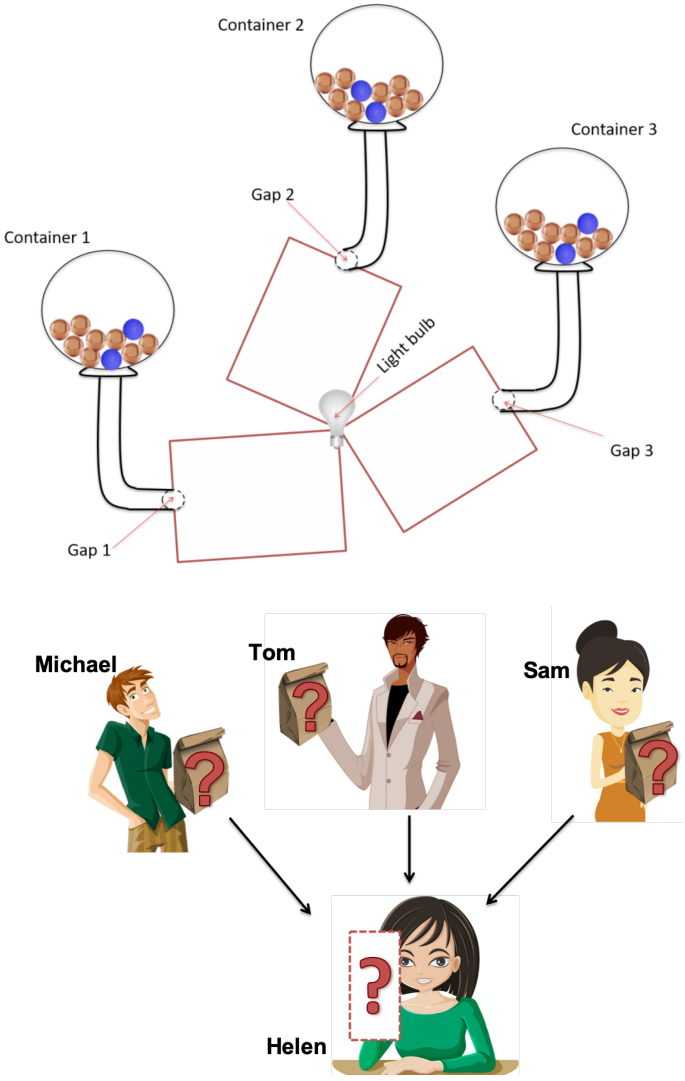


Figure 2.18: Graphical representations of the cover stories presented to participants in Experiment 3. The top image was featured in the balls and container cover story and the bottom one in the dinner party cover story.

Table 2.12: Inference types and questions found in the questionnaire for Experiment 2.

Question Num.	Inference Type	Key Inferences	Question Type
1	**Priors**	$P(C_1)$	Quantitative
2		$P(C_2)$	Quantitative
3		$P(C_3)$	Quantitative
4	**Independence**	$P(C_2 \mid C_1)$	Qualitative
5		$P(\sim C_3 \mid \sim C_2)$	Qualitative
6		$P(C_1 \mid \sim C_3)$	Qualitative
7 , 8	**Diagnostic Reasoning**	$P(C_1 \mid E)$-*R*-$P(C_1)$	Qual. & Quant.
9 , 10		$P(C_2 \mid E)$-*R*-$P(C_2)$	Qual. & Quant.
11 , 12		$P(C_3 \mid E)$-*R*-$P(C_3)$	Qual. & Quant.

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

vide 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. $P(C_2 \mid C_1)$ given no knowledge of whether E is present or not. To investigate participants' diagnostic reasoning I 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.

2.1.9.3 Results

Participants' answers to all qualitative in the inference questionnaire are represented in Figure 2.19 and their responses to all quantitative questions are in Figure 2.20.

Overall Performance As in Experiments 1 and 2, to test for a main effect of cover story and/or priors condition on participants' judgment accuracy throughout the inference questionnaire I initially 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 $\pm$ .02 of the normative probability estimate. This allowed for a comparative measure of participants' accuracy in both qualitative and quantitative types of inferences. Subsequently, I 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. The same

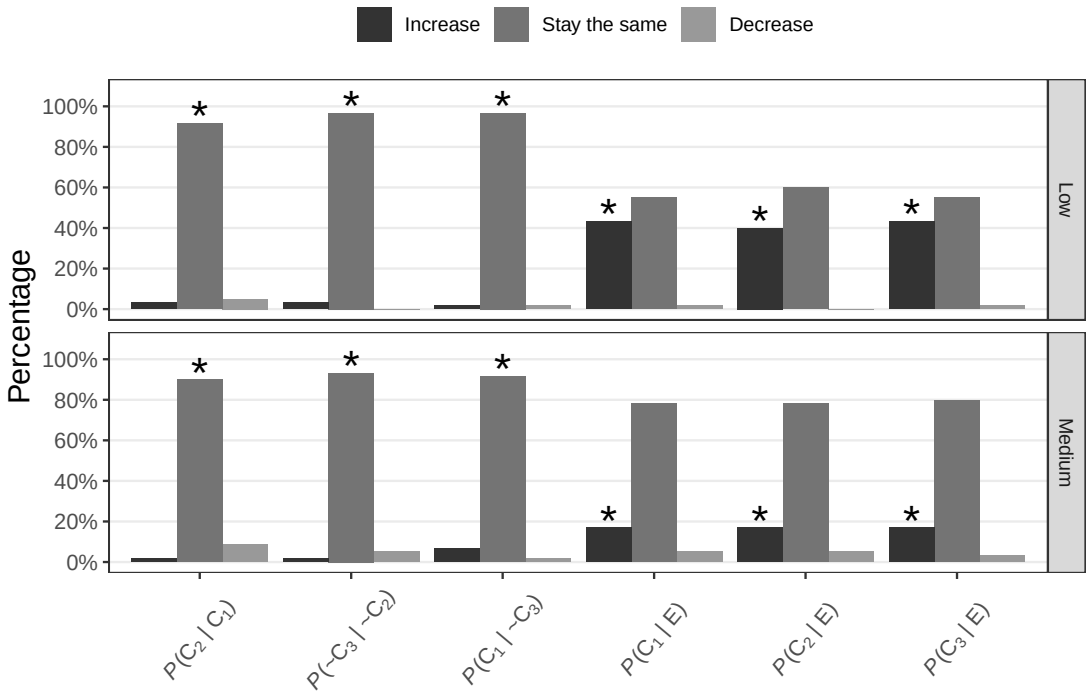


Figure 2.19: Distribution of participants' responses to qualitative questions in Experiment 3. Asterisks above the bars indicate normative answers.

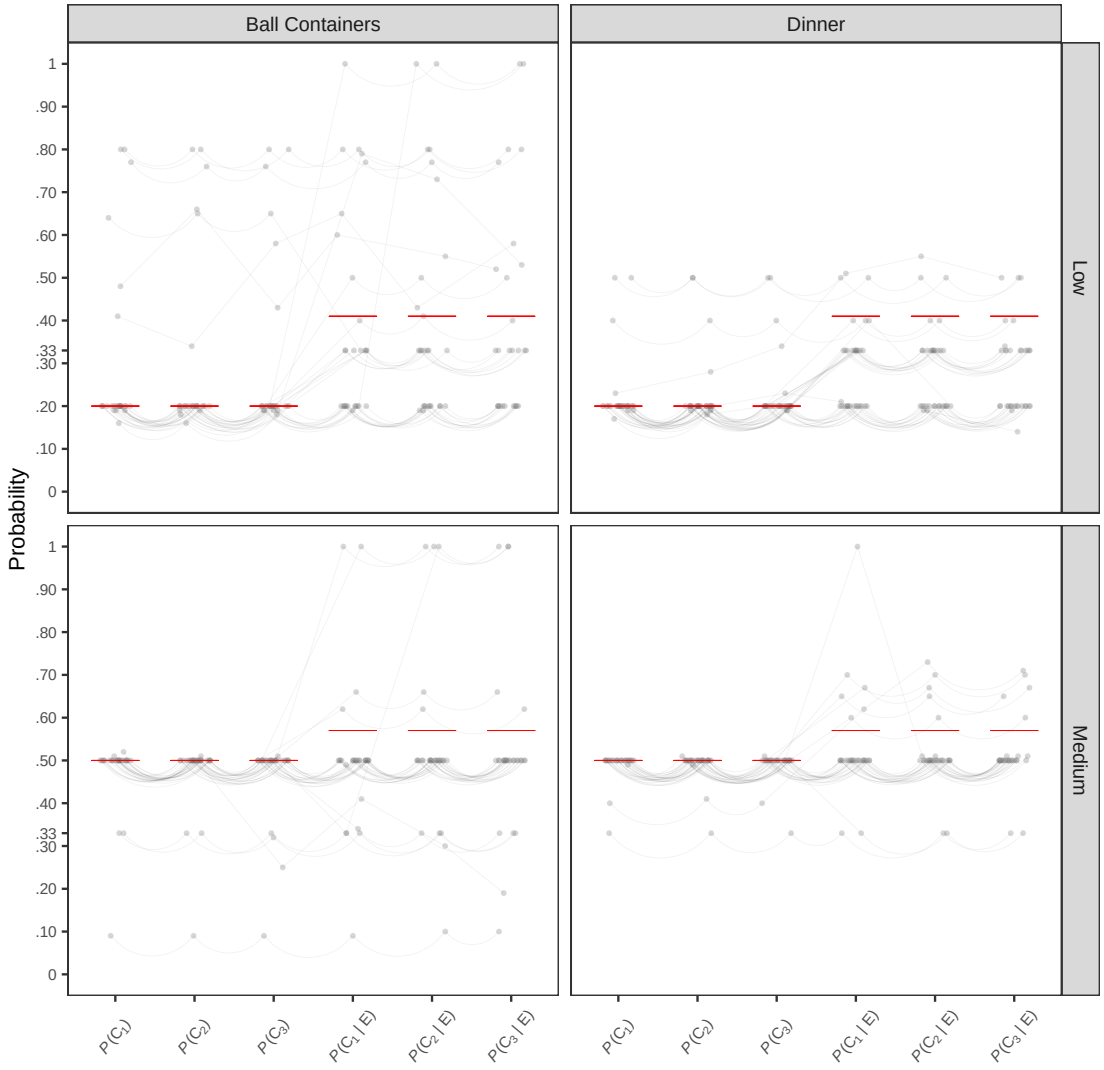


Figure 2.20: Participants' responses to quantitative questions in Experiment 3. Red horizontal lines are correct (normative) answers. Gray lines between data points depict how participants changed their probability estimates from one questions to another, with curved lines indicating that a participant did not change (within $\pm .02$) their probability estimate.

was done for the questions regarding independence and qualitative and quantitative diagnostic reasoning. This leaves four coded question-types regarding: priors, independence, qualitative diagnostic reasoning, and quantitative diagnostic reasoning.

To determine the effect of the manipulations on participants' overall performance throughout the task I 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). I found no main effect of Priors, $z = -0.89$, $p = .36$ and no main effect of Cover story, $z = -0.7$, $p = .49$. There was also 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 the predictors were centered, this implied that the data grand mean fits the data no worse than the model which includes both predictors.

Given there was no effect of scenario or priors on participants' performance, I 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. $P(C_1)$, $P(C_2)$ and $P(C_3)$.

Independence of causes For a breakdown of the frequency of participants' answers to qualitative independence questions see Figure 2.19. Collapsing across conditions, 89% of participants correctly answered *all three* questions

relating to independence (i.e. Qs 4, 5, and 6 in Table 2.12).

Diagnostic Reasoning For a breakdown of the frequency of participants' answers to qualitative diagnostic reasoning questions see Figure 2.19. In regards to diagnostic reasoning, 26% of participants correctly answered all three *qualitative* diagnostic reasoning questions (Qs 7, 9, and 11 in Table 2.12) and only 2.5% of participants correctly answered all three *quantitative* diagnostic reasoning questions (Qs 8, 10, and 12 in Table 2.12).

Diagnostic split In order to test the diagnostic split hypothesis I firstly collapsed the cover story condition and subsequently computed the proportion of participants who, having given the correct priors ($\pm$ .02) for all three causes ($N = 100$, or 84% of all data), updated the probabilities of $P(\mathrm{C}_1 \mid \mathrm{E})$, $P(\mathrm{C}_2 \mid \mathrm{E})$ and $P(\mathrm{C}_3 \mid \mathrm{E})$ to .33 ($\pm$ .02) each. This proportion was 34% in group reasoning with low priors and 3.8% in group reasoning with medium priors (of $N = 100$). A Chi-Square test of independence illustrated showed that these proportions significantly differed from each other, $\chi^2(1) = 13.48$, $p < .001$. The 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 Experiments 1 and 2 analyses, I collapsed all data and cross-tabulated responses of participants who correctly answered all three priors questions. Table 2.13 illustrates that the diagnostic split hypothesis accounted for about 18% of violations in quantitative diagnostic reasoning (of $N = 100$).

Propensity interpretation Although I 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 2.13 shows that about 67% of the participants who failed *quantitative* diagnostic reasoning reasoned in line with the propensity interpretation (i.e. they provided estimates $P(C_i \mid E) = P(C_i)$ ($\pm$ .02) for all three causes) (of $N = 100$). Table 2.14 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) (of $N = 100$). These results suggest that the propensity hypothesis accounted for a significant proportions of failures in diagnostic reasoning.

2.1.9.4 Discussion

Like in Experiments 1 and 2, in this experiment 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 Experiments 1 and 2 and suggest that participants had a good understanding of the causal structure, parameters, and the cover story they were reasoning with. Despite this, I 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 the diagnostic split hypothesis, I found that it accounted for about 18% of violations in diagnostic reasoning (of $N = 100$). More specifically, I found that a significant portion of participants employed this strategy in the

Table 2.13: 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 3. Total $N = 100$.

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

group reasoning with low priors, who increased their probabilities of $P(C_i)$ from .2 to .33. By contrast, this strategy was barely 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. The findings therefore strengthen the notion that the diagnostic split hypothesis is 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 I observed a dearth of participants who engaged in the diagnostic split strategy when rea-

Table 2.14: 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 3. Total $N = 100$.

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

soning 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 3 solidify the presence of the diagnostic split hypothesis (in the normative direction of update) and serve to demonstrate that underlying participants' updating behaviour in Experiments 1 and 2 (attributing .5 to each cause) was not a lack of confidence or an unawareness of the task, but rather engagement in a specific strategy.

Another updating behaviour that accounted for a large cluster of participants' data is encompassed by the propensity hypothesis. I 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 I have not explicitly tested the propensity hypothesis in Experiment 3

these proportions provide further empirical support for it.

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

2.1.10 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.

I carried out three 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, I explicitly stated the prior probabilities of the causes found in the model and re-elicited these from participants in order to ascertain that these were accepted. Moreover, I utilised relational qualitative and quantitative question formats to elicit probabilistic inferences from participants. This allowed for to the assessment of 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, I 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 all three experiments there was 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 the experiments and given encouraging finding from Rehder (2014a) who also employed a version of qualitative forced choice questions, utilising qualitative questions to address this understanding might be a promising way forward.

In addition, in Experiments 1 and 2 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 $P(E \mid \sim C_i, \sim C_j) = 0$ which implies that $P(C_i \mid E, \sim C_j) = 1$).

Despite the encouraging findings regarding priors, independence, and logic, the 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 Experiments 1 and 2 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 corresponded to different amounts of explaining away only in diagnostic reasoning. Notably however, the results suggested that the proportions of participants correctly answering the *qualitative* questions did directly correspond to the normative amount of explaining away, a finding that should further be explored. In addition, findings from both experiments suggest that deviations from the normative model observed in the 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.

2.1.10.1 Diagnostic split

The findings of the three experiments suggest that some people do split the probability space between the causes when engaging in diagnostic reasoning. As such, I found that a significant proportion of participants' answers aligned

with predictions made by the diagnostic split hypothesis. Furthermore, Experiment 2 explored diagnostic reasoning with unequal priors. The findings in this experiment suggested the participants who engaged in diagnostic split reasoning divided the probability space between the causes in different ways: some have provided equal probability to both causes in diagnostic reasoning and some have reflected the ratio of the priors of the causes and divided the probability space between the two causes according to that ratio in diagnostic reasoning. This suggests that there are at least two ways that drive the way participants who engage in diagnostic split reasoning divide the probability space: (i) split it equally among all causes and (ii) split the probability space following the ratio of the prior probability of causes. Experiment 3 tested the strategy in the context of three causes and excluded an alternative explanation of the findings from Experiments 1 and 2 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 3, the diagnostic split hypothesis needs to be modified 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 $P(\mathrm{C_1} \mid \mathrm{E}) + \cdots + P(\mathrm{C_i} \mid \mathrm{E}) + \cdots + P(\mathrm{C_n} \mid \mathrm{E}) = 1$ when the set-up is deterministic and $P(\mathrm{C_i}) \leq \frac{1}{n}$.

Crucially, through the use of cross-tabulations I was able to illustrate that adopting a diagnostic split strategy accounted for 51% of observed deviations in Experiment 1 (of $N = 386$, or of 85.2% of all data) and 47% in Experiment 2 (of $N = 208$, or of 79.1% of all data) in both quantitative diagnostic reasoning and quantitative relational explaining away. In Experiment 3 approximately 18% of violations in quantitative diagnostic reasoning could be attributed to a diagnostic split strategy (of $N = 100$, or of 84% of all data). Ultimately, this suggests that this strategy contributes significantly to the observed violations of explaining away.

So far I 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.

2.1.10.2 Propensity interpretation

The findings from the three experiments also suggest that a large number of participants remained at the priors when answering diagnostic reasoning and

direct explaining away questions. Moreover, Experiments 1 and 2 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 I 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, the propensity hypothesis was able to account for a significant amount of insufficiency in explaining away. The cross-tabulations showed that the propensity interpretation was able to account for 53% of violations in Experiment 1 (of $N = 386$, or of 85.2% of all data) and 27% in Experiment 2 (of $N = 208$, or of 79.1% of all data) in both the quantitative diagnostic reasoning and quantitative relational explaining away; also over 70% violations in Experiment 1 and over 54% violations in Experiment 2 in qualitative diagnostic reasoning, direct and relational explaining away could also be explained by the propensity hypothesis. These percentages support the hypothesis 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 ex-

plain 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 I expect the propensity interpretation (or anchoring at the base rate) to 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 be more prone to people neglecting the priors. These should be explored in the future work.

Finally, I 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, problems for propensity interpretation have been raised, such the Humphreys' paradox, that challenge the idea that it can be reconciled with the axioms of probability which are the bedrock 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 Bayesian

updating (i.e. 'conditionalization') which means 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 2.1.6.2 I 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 thesis to argue for or against the normative status of the propensity interpretation. I simply find that the propensity interpretation is a good descriptive account of the findings on explaining away.

2.1.10.3 Limitations

A few important limitations are in order. First, in all three experiments priors and conditional probabilities were communicated textually and graphically to participants. I have not explored whether the 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, I expect the propensity interpretation to be less pronounced. As a consequence I would expect less participants to stay at the priors in diagnostic reasoning and explaining away compared to the findings in the current study. However, I 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. There are, however, even further ways to communicate probabilities to participants apart from textual, graphical representations, and experiential learning include. For instance, one could employ summary descriptions and labelling. These all should be ex-

plored in future studies.

Second, I have only considered explaining away in a deterministic set-up. Admittedly this is fairly limiting from a perspective of the ecological validity of the findings. I proposed further avenues of research with respect to this limitation and have also argued that I expect to find similar results with respect to the diagnostic split hypothesis even in non-deterministic set-ups. The propensity interpretation seems to be particularly pronounced in deterministic set-up and it may be less pronounced in non-deterministic ones. See Experiments 4 and 5 for findings regarding the two hypothesis in a non-deterministic set-up.

Third, in the experiments I 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 I have not used, for instance, the frequency format response scale is that (i) given the events in the 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, I believe, steer them away from the propensity probability

interpretation towards the frequency interpretation (which is out of the scope of this thesis) thus reducing the power of the 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, I 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).

2.2 Extending explaining away: Learning (about) multiple pieces of evidence[27]

In the first part of this chapter I have discussed explaining away situations where there are two independent causes and one common effect. We are, however, often exposed to more complex situations where there are multiple pieces of evidence that could be accounted by multiple causes. For example, a person is no more or less likely to have a bacterial infection if they have a viral infection, but both a viral infection and a bacterial infection can cause a rash and also a swelling. These situations also give rise to explaining away reasoning, but have a number of features that do not occur in explaining away situations with only one effect. For instance, learning that one effect has happened will make the causes dependent. This is also true in the explaining away situations with one effect. However, because the two causes are now probabilistically dependent, learning that another effect has also happened will have a different impact on the probabilities of the two causes then if the causes were probabilistically independent. This entails that in explaining away situations, every other additional piece of evidence will result in (i) these pieces of evidence having different impact on the causes from a CBN modeling perspective because the causes are no longer independent and (ii) the normative amount of explaining away would be different (and most likely it will be reduced if the causes become positively correlated) then in typical explaining away situations with independent causes.

The second feature of reasoning with multiple pieces of evidence relates to

[27]This section is based on work from Tešić and Hahn (2019).

its sequential character. In real world contexts of reasoning about evidence, that evidence frequently arrives sequentially. The standard normative CBN framework for probabilistic reasoning yields the same ultimate outcome whether multiple pieces of evidence are acquired in sequence or all at once, and it is insensitive to the order in which that evidence is acquired. From an empirical perspective, it is then interesting to explore different orders in which evidence becomes available. The third feature relates to learning a new piece of evidence whose potential existence we were not even aware of. We often cannot anticipate in advance what kinds of evidence we will eventually encounter. For example, we may take our car to the mechanic because it started making a noise that we believe is concerning. The mechanics may inform us that there is a problem with the engine's wastegate, a part of our car's engine that we were not aware it even existed. This raises the question of what we do to our existing models when we encounter new variables to consider. However, little is known about what happens when evidence sets are expanded incrementally, both from the normative CBN perspective and a descriptive perspective.

In this part of the chapter I discuss explaining away situations where one is required to reason with multiple pieces of evidence. More specifically, I discuss (i) how to model from a CBN perspective the situations when one is made aware of a potential existence of a piece of evidence that one was not aware before; (ii) how people reason in these kinds of situations when the evidence is presented sequentially (in different orders) vs. all evidence is presented at the same time, and where different pieces of evidence have different diagnosticity (i.e. have different impacts on the causes). In this part I, thus, explore situations that are not characterised by a deterministic set-up like in the previous part

(i.e. situations where the presence of at least one cause entails the presence of an effect). This will allow to extend some of the findings from the previous part to the non-deterministic situations and would arguably increase the ecological validity of these findings.

I explore both (i) and (ii) in a segment of (relational) explaining away, namely diagnostic reasoning. The primary reason for not exploring all segments of explaining away reasoning in this part were the findings from Liefgreen et al. (2018) which suggested that people find (direct) explaining away reasoning in situations with multiple causes and multiple effect sufficiently difficult that they tend to disengage. As I have adopted a very similar experimental paradigm as used by Liefgreen et al. (2018), that I restricted the exploration of (i) and (ii) to diagnostic reasoning. Nonetheless, the first three experiments have suggested that people's accuracy on diagnostic reasoning is indicative of their accuracy in (direct) explaining away: participants who correctly answered (qualitative) diagnostic reasoning questions were more likely to answer (direct) explaining away questions. Thus, even though I do not explore (direct) explaining away in the following experiments, the findings could still be indicative of people's performance on (direct) explaining away questions.

2.2.1 Introduction

Imagine the following situation. Tom wakes up one morning and notices a rash on his skin. He does not think the rash is a big deal, but after a couple of days the rash is still present so he decides to see a doctor. Before he visits a doctor he thinks that the rash is either caused by a bacterial or a viral infection

or, perhaps, both. Tom also believes that having a bacterial infection does not make one more or less likely to have a viral one and vice versa; in other words, Tom thinks that the two types of infection are independent. The doctor agrees with him that the rash could be caused by a bacterial and/or a viral infection and that the two types infections are independent. However, she additionally informs Tom that he also has a swelling he didn't notice, which can also be caused by a bacterial and/or a viral infection. Furthermore, she tells him that either type of infection is more likely to cause the swelling than the rash. How do (should) Tom and the doctor revise their beliefs about multiple independent causes given multiple pieces of evidence of different diagnosticity?

From a normative standpoint, many would argue that the answer is encoded in a CBN with 2 common effects and 2 independent causes.[28] For instance, the CBN in Figure 2.21 would model the situation described above: C_1 = viral infection, C_2 = bacterial infection, E_1 = rash, and E_2 = swelling.

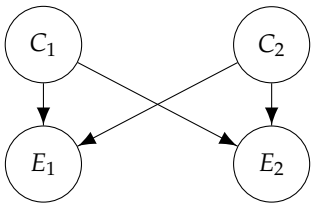


Figure 2.21: CBN with 2 independent causes and 2 common effects.

To fully parameterize CBN from Figure 2.21, one needs to specify the following probabilities:

$$P_1(\mathrm{C}_1) = c_1 \quad , \quad P_1(\mathrm{C}_2) = c_2$$

[28] B. K. Hayes, Hawkins, Newell, Pasqualino, and Rehder (2014) have used a dynamic CBN to model these kinds of situations. However, in this part of the chapter I employ static CBNs as there are no significant differences in the formalism in this case.

$$\begin{aligned} P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \mathrm{C}_2) &= \alpha_1 \quad , \quad P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = \beta_1 \\ P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \mathrm{C}_2) &= \gamma_1 \quad , \quad P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = \delta_1 \\ P_1(\mathrm{E}_2 \mid \mathrm{C}_1, \mathrm{C}_2) &= \alpha_2 \quad , \quad P_1(\mathrm{E}_2 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = \beta_2 \\ P_1(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \mathrm{C}_2) &= \gamma_2 \quad , \quad P_1(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = \delta_2 \end{aligned} \tag{2.4}$$

$P_1(\mathrm{C}_1)$ and $P_1(\mathrm{C}_2)$ are the prior probability of the two causes and the remaining probabilities are a part of the conditional probabilities tables for the two effects. The doctor then could use this CBN to, via diagnostic reasoning, update her beliefs about the probability that Tom has a viral infection after learning that Tom has a rash by calculating $P_1(\mathrm{C}_1 \mid \mathrm{E}_1)$. After additionally learning that Tom also has swelling the doctor could further update her probability of Tom having a viral infection by calculating $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2)$ (similarly for the bacterial infection).

However, it is somewhat accidental that Tom first noticed the rash and not the swelling. He could have plausibly first seen the swelling and gone to the doctor and then noticed the rash. Would the CBN calculation be different in this scenario? It depends. If the rash and the swelling are not equally diagnostic of the two causes as is suggested by the example, then it is possible that $P_1(\mathrm{C}_1 \mid \mathrm{E}_1) \neq P_1(\mathrm{C}_1 \mid \mathrm{E}_2)$, in which case the doctor's degrees of belief about a viral infection after first learning that Tom has swelling would not be equal to those where she first leaned about the rash. However, after learning the second effect the order in which the effect appear no longer matters: that is, $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2)$ is always equal to $P_1(\mathrm{C}_1 \mid \mathrm{E}_2, \mathrm{E}_1)$. Moreover, from a formal point of view the probability of a case after learning the two pieces of evidence in a sequence and updating the probability of the cause after each piece of evidence was known is equal to the probability of the cause that was updated from the

prior probability after learning the two pieces of evidence at the same time.

It is then interesting from an empirical point of view to investigate whether people are sensitive to these different ways the evidence was made available and whether they update the causes differently depending on the order in which the effects appear. Studies on sequential diagnostic reasoning have sought to tackle exactly these issues (see Meder & Mayrhofer, 2017b; Hogarth & Einhorn, 1992). They presented participants with a sequence of effects and asked them to reason from multiple effects to causes either with each effect they learned (step-by-step procedure) or after they learned about the whole sequence of effects (end-of-sequence procedure) (see Hogarth & Einhorn, 1992; Rebitschek, Bocklisch, Scholz, Krems, & Jahn, 2015). Their studies were primarily interested in investigating primacy effects (whereby most of the evidential weight is given to the first piece of evidence) and recency effects (whereby most of the evidential weight is given to the most recent pieces of evidence). Meder and Mayrhofer (2017a) investigated sequential diagnostic reasoning by providing participants with verbal information regarding the strengths of the causes instead of a more quantitative information (like the CPTs) and found that participants are remarkably accurate in their judgements. However, all these studies investigated only situations where the causes were mutually exclusive and exhaustive causes. These situations are not explaining away situations and would be modeled with one node for all causes (see Figure A.1). An exception is a study by B. K. Hayes et al. (2014). They have investigated a scenario where two symptoms could be produced by two independent causes. However, in their study both effects had exactly the same diagnosticity (i.e. the same CPT) and for that reason there were no order effects, i.e. it did not matter whether

one learnt first E_1 or E_2, $P_1(C_1 \mid E_1) = P_1(C_1 \mid E_2)$.

One of the goals of this part of the chapter is to empirically investigate people's ability to reason diagnostically from multiple effects with different diagnosticities (CPTs) to multiple independent causes. More specifically, the aim was to test how people's judgements compare to the normative answer from CBNs such as the one in Figure 2.21 by manipulating the way in which multiple pieces of the evidence of different diagnosticity are presented (in a particular order or at the same time) and the way judgements about the causes are elicited from the participants (step-by-step (SbS) or all-at-once (AaO)).

Another interesting issue emerges when reasoning with independent causes. Not only can we learn the evidence sequentially, but we can sequentially learn about new pieces of evidence that we were not previously aware of that may also influence our beliefs about the causes. In technical parlance, we may need to expand the algebra to incorporate the new variables. Consider Tom from our example. Initially Tom only knew about his rash and, based on that knowledge, he updated his probabilities of the two causes. Unlike the doctor, Tom did not even know that the two types of infection could also cause swelling. It is only after he visited his doctor that he learned about the another potential effect and the occurrence of that effect. At the time he only knew about the rash he updated the probabilities of the two causes on the basis of a CBN model with only three nodes: two independent causes and one common effect while the doctor always had in mind the CBN from Figure 2.21. Despite operating with two different CBNs, both Tom and the doctor would arrive at the same probabilities (assuming the same priors and CPTs for the effect) at this first step. The next step is, however, crucial. After learning about the swelling,

the doctor would simply learn that the new piece of evidence has occurred and update the probabilities of the causes based on the CBNs from Figure 2.21. Tom, by contrast, might do one of the at least two following possibilities: (1) forget about his original 3-node network and create a new 4-node one like the one in Figure 2.21 in which case he would arrive at the same estimates as the doctor; or (2) take his (and doctors) previous estimates of the two causes based on only one piece of evidence and have them as new priors in his new 3-node network with the second piece of evidence as a common effect (see Figure 2.22). In the latter case he would be 'splitting' the CBNs from Figure 2.21 into two 3-node networks.

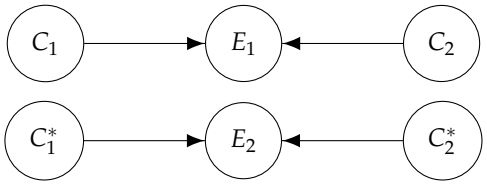


Figure 2.22: 'Split' CBN from E_1 to E_2

$$
\begin{aligned}
P_1(\mathrm{C}_1) = e_1 \quad &, \quad P_1(\mathrm{C}_2) = e_2 \\
P_2(\mathrm{C}_1^*) = P_1(\mathrm{C}_1 \mid \mathrm{E}_1) \quad &, \quad P_2(\mathrm{C}_2^*) = P_1(\mathrm{C}_2 \mid \mathrm{E}_1) \\
P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \mathrm{C}_2) = \alpha_1 \quad &, \quad P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = \beta_1 \\
P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = \gamma_1 \quad &, \quad P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = \delta_1 \\
P_2(\mathrm{E}_2 \mid \mathrm{C}_1^*, \mathrm{C}_2^*) = \alpha_2 \quad &, \quad P_2(\mathrm{E}_2 \mid \mathrm{C}_1^*, \neg\mathrm{C}_2^*) = \beta_2 \\
P_2(\mathrm{E}_2 \mid \neg\mathrm{C}_1^*, \mathrm{C}_2^*) = \gamma_2 \quad &, \quad P_2(\mathrm{E}_2 \mid \neg\mathrm{C}_1^*, \neg\mathrm{C}_2^*) = \delta_2
\end{aligned}
\tag{2.5}
$$

Eq. (2.5) specify the priors and the CPTs of the two networks. Although one might intuitively think that Tom will arrive at the same probabilities as the doctor even in the case where he models the situation as in Figure 2.22, that turns

out to be true only under very specific conditions, some of which may violate common assumptions in causal Bayesian reasoning (see Appendix A.3). Less technically, this is because once one learns evidence (E_1) and updates the probabilities of the two causes (C_1 and C_2) in a common-effect CBN, the two previously independent causes become dependent: although $P_1(\mathrm{C}_1 \mid \mathrm{C}_2) = P_1(\mathrm{C}_1)$, generally $P_1(\mathrm{C}_1 \mid \mathrm{C}_2, \mathrm{E}_1) \neq P_1(\mathrm{C}_1 \mid \mathrm{E}_1)$ as discussed in the first part of this chapter. This dependency is preserved in the full CBN network in Figure 2.21 even *before* one learns the second piece of evidence (E_2) and again updates the probabilities of the two causes. However, in the lower 3-node CBN in Figure 2.22 the dependency is lost since it is assumed that C_1^* and C_2^* are independent *before* observing E_2. Even though $P_2(\mathrm{C}_1^*)$ is equal to $P_1(\mathrm{C}_1 \mid \mathrm{E})$ and $P_2(\mathrm{C}_2^*)$ is equal to $P_1(\mathrm{C}_2 \mid \mathrm{E})$, making C_1^* and C_2^* independent and not preserving the independence induced by the first piece of evidence will result in final probability estimates of the two causes, i.e. their estimates *after* learning both pieces of evidence, to most likely diverge on the two different modeling strategies. More specifically, the final estimates of the two causes will always be higher according to the 'split' CBN in Figure 2.22 than those according to the full one in Figure 2.21 precisely because the full one accounts for the above-mentioned dependency and the 'split' one does not. Moreover, when the diagnosticity of the two pieces of evidence is different (as is the case in this study), the height of the final estimates in the 'split' CBN will depend on the order the evidence is observed: learning E_1 then E_2 will result in the final estimates of the causes that are different to those that result from learning E_2 then E_1 (as previously mentioned, whether we learn E_1 first then E_2 or vice versa does not affect the final probability estimates of the causes in the full CBN). It is also worth point-

ing out that this divergence only happens when the causes are independent. If the causes are mutually exclusive and exhaustive, one can safely 'split' the full network into multiple ones without worrying about ending up with different estimates (see Appendix A.4).

It is worth pointing out that from a standard CBN perspective, there is not a right or wrong way of modeling situations where one needs to incorporate new variables in their already existing CBN models. The CBN theory is silent when it comes to extending algebra and defining a new probability distribution over a new algebra. The two modeling strategies from above are chosen mostly because they seem to be plausible ways of incorporating a new variable in the specific context of adding a new common effect variable to an already existing structure with two independent causes and one common effect. In situations where one would need to add a new common cause to a common effect structure (for example, adding a new common cause to C_1 and E_1 in the upper 3-node CBN from Figure 2.22) would presumably be more difficult to do from a CBN modeling perspective as it would require redefining some of probability distributions that have already been defined in the old causal structure (for example, it would require expressing the prior probability of C_1 in terms of conditional probabilities as in the new causal structure C_1 would be dependent on the common cause and it would require redefining the conditional probability table for E_1).

To address the issue in a more principled way one could try to define a similarly measure between the causal structures and use this measure to decide how to incorporate a new variable and define a probability distribution over the new algebra. The work on these kinds of similarity measures is still in

its infancy. For example, Eva, Stern, and Hartmann (2019) introduce multiple similarity metrics and ways to prioritize these metrics to measure the similarity between two causal graphs. However, their work is limited to comparing two causal graphs that differ in their causal structure (i.e. the way the variables are connected or not connected to each other), but not in their algebra as the two graphs had to include the same variables. This part of the chapter is thus contributing in a limited way to understating two different modeling strategies when it comes to including new variables that were not part of previous model's algebra (see Appendices A.3 and A.4).

From an empirical point of view, to the best of my knowledge no study has yet investigated sequential diagnostic reasoning with sequentially learning the algebra. In the literature mentioned above participants were presented with all the variables and the causal/probabilistic information related to them before they started making judgements about the causes. Even in such contexts, it is worth looking at order effects because it has long been recognized that order effects may be particularly diagnostic with respect to the processes underlying the formation of a judgment. Specifically, there is a long literature concerned with order effects in contexts such as impression formation (N. H. Anderson, 1965) or numerical estimation (Jacowitz & Kahneman, 1995). However, the investigation in this chapter goes beyond this as one of the aims of this part is to examine how reasoners fare in probabilistic reasoning contexts where they are faced with entirely new variables. This issue has, to the best of my knowledge, not been explored. In many scientific and everyday situations we must make judgements about potential causes given effects without being aware of other potential effects that could also inform our judgements. Thus, the main aim of

the study presented below was to examine how people reason with multiple pieces of evidence when they successively learn not just that some piece of evidence obtains, but also that there is another potential piece of evidence not known before. I compared participants' estimates to both the full network's predictions (Figure 2.21) and the 'split' networks' predictions (Figure 2.22).

2.2.2 Experiment 4

In this experiment I investigated the influence of manipulating algebra and evidence learning on probabilistic diagnostic reasoning judgements of the two independent causes. Participants were prompted to reason with either the full 4-node model (Figure 2.21) from the outset or they learned in stages that there is another possible effect of the two causes. Further, participants either observed the effects in one of the two sequences or they observed both effects at once. The prior probabilities of the cases and CPTs of the effects were the same in all conditions: $P(\mathrm{C}_1) = P(\mathrm{C}_2) = .15$; $P(\mathrm{E}_1 \mid \mathrm{C}_1, \mathrm{C}_2) = .99$, $P(\mathrm{E}_1 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = P(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = .7$, $P(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = 0$; $P(\mathrm{E}_2 \mid \mathrm{C}_1, \mathrm{C}_2) = .6$, $P(\mathrm{E}_2 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = P(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = .2$, $P(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = 0$. For simplicity the priors of the causes were the same and the CPTs of the effects reflected different diagnosticities of the two effects.

2.2.2.1 Methods

Participants and Design A total of 271 participants (N_{MALE} = 101, M_{AGE} = 32.1 years; one participant identified as neither male nor female) were recruited from Prolific Academic (www.prolific.ac). All participants were native English speakers who gave informed consent and were paid £1.25 for taking part

in the present study, which took on average 13.9 minutes to complete. Participants were randomly assigned to one of the 2 (algebra: full or sequential) $\times$ 3 (evidence learning: all-at-once (AaO), step-by-step from E_1 to E_2 (SbS1), or step-by-step from E_2 to E_1 (SbS2)) = 6 between-participants groups (one group with 44 participants, 3 groups with 45 participants, and 2 groups with 46 participants).

Materials All participants were given the same cover story wherein rain (C_1) and a lawn sprinkler (C_2) (two binary and independent variables) could cause a wet lawn (E_1) and/or a wet exterior house wall (E_2) (a version of the cover story can be found in Pearl, 1988). The participants in the AaO condition completed an online inference questionnaire comprising of 10 comprehension questions (2 about the priors of the causes and 8 about the CPTs) and 2 test questions (one relating to $P(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2)$ and one to $P(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2)$). Everyone else completed the same 10 comprehension questions and 4 test questions (relating to $P(\mathrm{C}_1 \mid \mathrm{E}_i)$, $P(\mathrm{C}_2 \mid \mathrm{E}_i)$, $P(\mathrm{C}_1 \mid \mathrm{E}_i, \mathrm{E}_j)$, and $P(\mathrm{C}_2 \mid \mathrm{E}_i, \mathrm{E}_j)$). For the full materials used in Experiment 4 see Appendix A.5.

Procedure After giving an informed consent and basic demographic information, participants were shown the following instructions:

WELCOME!

You will now be presented with a situation and required to answer some questions related to the situation. Please make sure you read all the information carefully before answering the questions.

Throughout the survey you will be able to navigate forward and

backward by clicking on the appropriate buttons.

After reading this information, the participants in the full algebra condition were presented with a causal cover story (both in a textual and a visual form) which explained the relations between variables and probabilistic information relating to the priors of both causes (priors were textually communicated as a percentage chance). They were then asked 2 priors comprehension questions where they were required to restate the priors communicated to them in the cover story. Following that, participants were told the CPTs of the two pieces of evidence (also textually communicated as a percentage chance) and subsequently asked 8 comprehension questions regarding the CPTs (in a random order) where they were required to restate the CPTs. After completing the comprehension questions, participants in the AaO condition learned that both pieces of evidence occurred and were prompted to answer 2 test questions (one for each cause) presented in the same order. Participants in the SbS conditions first learned about one piece of evidence and answered 2 test questions relating to the 2 causes and then learned that the second piece of evidence occurred and asked final 2 questions. When answering the test questions participants were reminded of the priors of the causes and the CPTs of each piece of evidence, as well as their previous estimates of the two causes (in the SbS conditions).

Participants in the sequential algebra condition were initially told a cover story (both in a textual and a visual form) including only two causes and one effect. As in the full algebra condition, they were told the priors of the causes (percentage chance) and asked 2 priors comprehension questions. In contrast to the full algebra contention, they were then told CPTs (percentage chance) regarding only one piece of evidence and completed 4 comprehension questions

related to CPTs (in both the AaO and the SbS conditions). This was followed by 2 test questions relating to the probability of the causes given that one piece of evidence was observed (only in the SbS conditions). Participants then additionally learned that there is another piece of evidence potentially relevant to the probability estimates of the two causes. They learned the CPTs for the second piece of evidence and completed 4 comprehension questions followed by 2 test questions prompting them to estimate their confidence in the causes happening given the additional piece of evidence obtained. Again, participants were reminded of the priors of the causes, CPTs (but only for the current piece of evidence), and their previous estimates of the two causes (in the SbS conditions). In the AaO, after completing the first 4 comprehension questions participants were not told that the evidence obtained. Rather, they went on to learn that there is another potentially relevant piece of evidence, completed additional 4 comprehension questions, and were subsequently then told that both pieces of evidence obtained. After that, participants were reminded of the priors, CPTs (for the both pieces of evidence) and completed 2 test questions regarding the probabilities of the two causes.

In all conditions the test questions prompted participants to provide percentage confidence (0–100%) of C_i given one or two effects. For example, after learning that E_1 occurred, they were asked (in SbS1 condition) a diagnostic reasoning questions: "How confident are you that it **rained** overnight now that you know that the lawn is wet?" After additionally learning E_2 occurred they were asked: "How confident are you that it **rained** overnight now that you know that *both* the lawn *and* the house wall are wet?" (the full algebra condition) or "How confident are you that it **rained** overnight now that you know

that the house wall is also wet?" (the sequential algebra condition). All participants provided explanations for each answer to the test questions.

2.2.2.2 Results

Comprehension To find out the proportion of participants who accepted the priors and the CPTs, I computed the numbers of participants who answered all 10 comprehension questions correctly ($\pm.02$). Out of 271 participants, only 61 answered all 10 comprehension questions correctly: in the full algebra condition, $\text{Group}_{\text{AaO}} = 8$, $\text{Group}_{\text{SbS1}} = 7$, and $\text{Group}_{\text{SbS2}} = 14$; in the sequential algebra condition, $\text{Group}_{\text{AaO}} = 13$, $\text{Group}_{\text{SbS1}} = 7$, and $\text{Group}_{\text{SbS2}} = 12$. Visually comparing the test questions estimates of the participants who correctly answered all 10 comprehension question (Figure 2.23) to those of the participants who incorrectly answered at least one of the 10 comprehension questions (Figure 2.24) seems to suggest that there was not much difference in the general trends and the ways participants answered the test question in these two groups.

A linear mixed effects model on participants' estimates to the test questions seems to support this view. The model had one fixed effect Comprehension (whether or not the participant answered all 10 comprehension questions correctly), a random intercept for each participant, and a random slope and a random intercept for each of the test questions. According to the model, there was no main effect of Comprehension ($t = .85$, $p = .41$). As there was no main effect of Comprehension, the following analysis was preformed on the full data set including all 271 participants. Figure 2.25 shows the responses of all 271 participants to the priors and test questions.

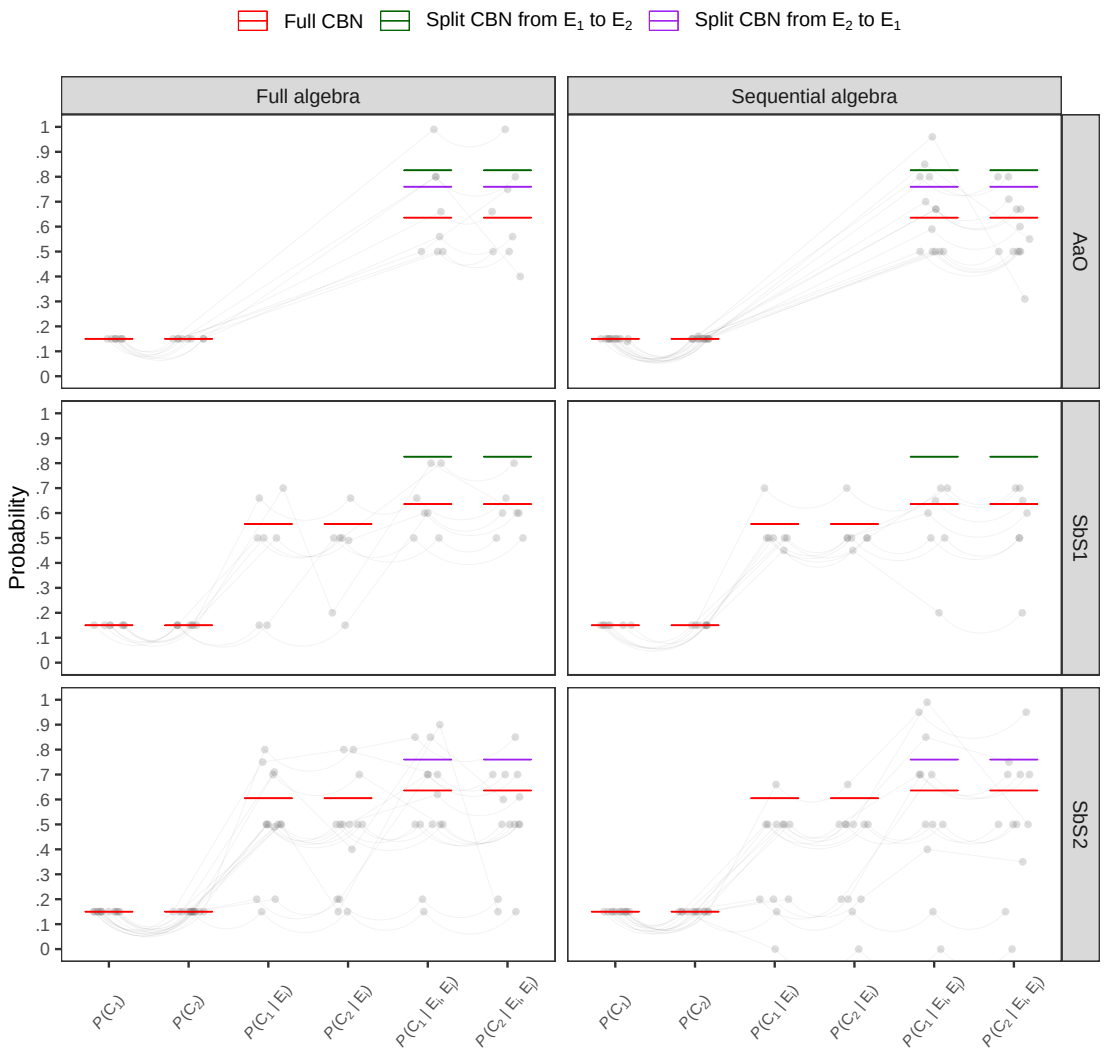


Figure 2.23: Responses of the participants who answered all 10 comprehension questions correctly to priors and test questions in Experiment 4. Red horizontal lines are correct (normative) answers according to the full BN model. Green and purple horizontal lines correspond to the predictions of the split CBN model. 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 $\pm$ .02) their probability estimate.

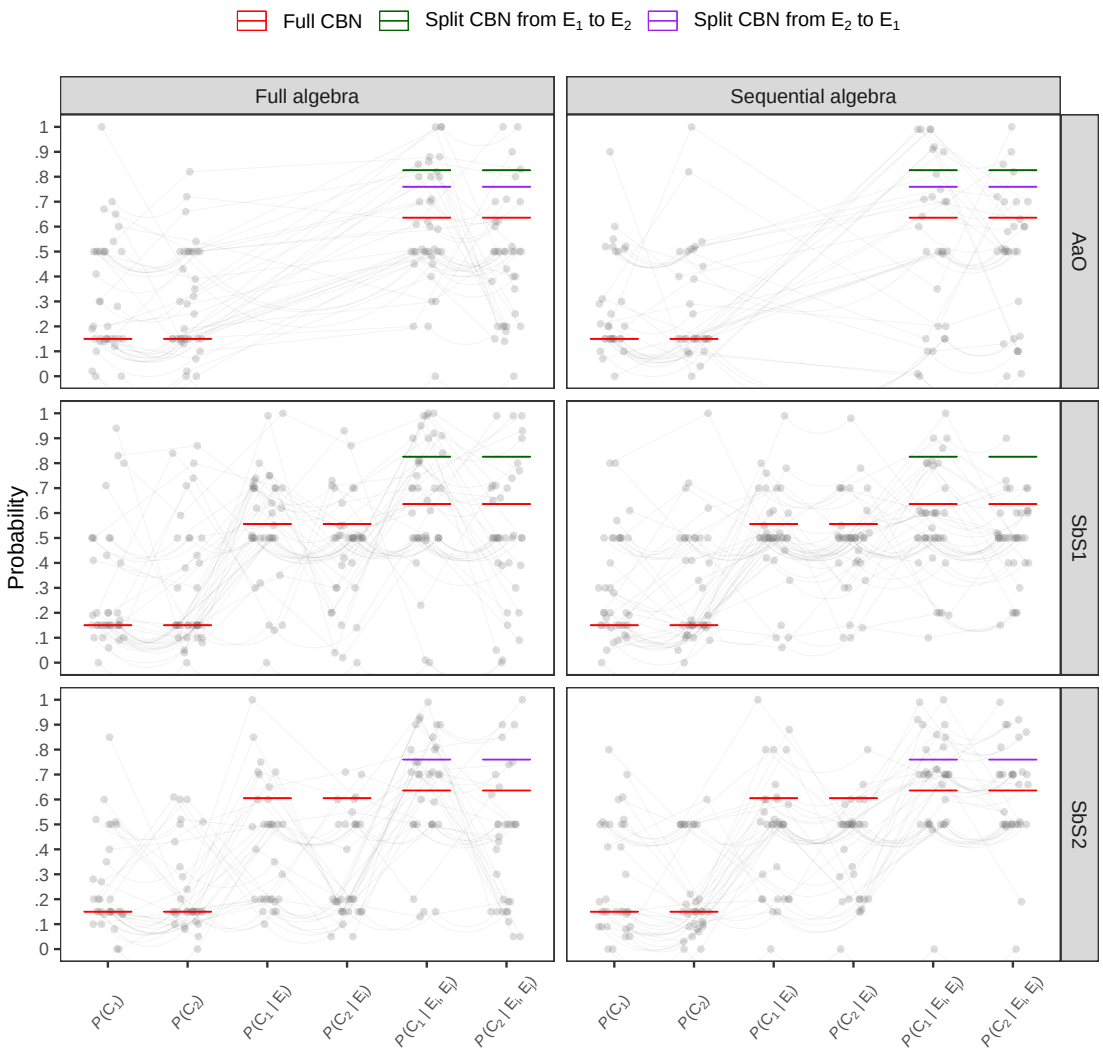


Figure 2.24: Responses of the participants who answered some of the comprehension questions incorrectly to priors and test questions in Experiment 4. Red horizontal lines are correct (normative) answers according to the full BN model. Green and purple horizontal lines correspond to the predictions of the split CBN model. 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 $\pm$ .02) their probability estimate.

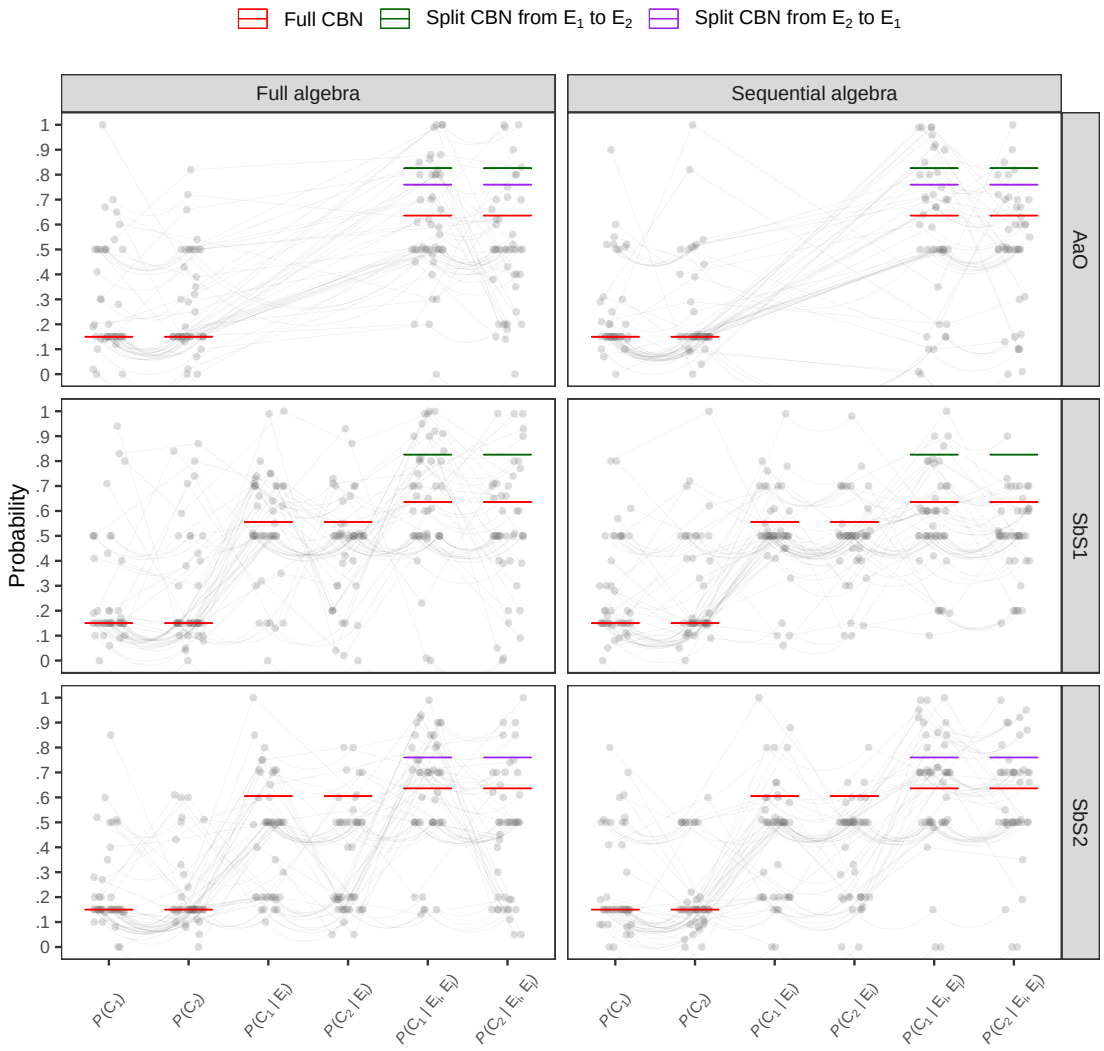


Figure 2.25: Responses of all 271 participants to priors and test questions in Experiment 4. Red horizontal lines are correct (normative) answers according to the full BN model. Green and purple horizontal lines correspond to the predictions of the split CBN model. Gray lines between data points depict how participants changed their probability estimates from one questions to another, with curved lines indicating that a participant did not change (within $\pm$ .02) their probability estimate.

Test questions To test the effect of the algebra and the evidence learning conditions on participants' estimates on the test questions, a linear mixed effects model was built. The model had two fixed effects, Algebra and Evidence learning, with a random intercept for each participant (there was no random slope for participant since algebra and evidence learning conditions vary between participants). I found a main effect of Evidence learning, but no main effect of Algebra (see Table 2.15). I also found no interaction between Algebra or Evidence learning. However, likelihood ratio tests showed that including the predictors in the model does not improve model fit compared to just having an intercept as a predictor ($\chi^2(3) = 6.11$, $p = .11$). That is, the data grand mean fits the data no worse than the model which includes both predictors.

Table 2.15: Linear mixed effect model results for test questions in Experiment 4
A=Algebra; EL=Evidence learning

	Estimate	95% CI	t-value	p
A	-6.51	[-17.76, 4.73]	-1.13	.26
EL	-0.53	[-1.03, -0.03]	-2.1	.04*
A$\times$ EL	3.28	[-17.76, 4.73]	1.29	.2

A finer grained analyses on the data within each group showed a significant difference between $P(C_1 \mid E_i)$ and $P(C_1 \mid E_i, E_j)$ in the full algebra SbS1 condition ($t(44) = -4.04$, $p < .001$); in the full algebra SbS2 condition both between $P(C_1 \mid E_i)$ and $P(C_1 \mid E_i, E_j)$ ($t(45) = -4.87$, $p < .001$) and $P(C_2 \mid E_i)$ and $P(C_2 \mid E_i, E_j)$ ($t(45) = -2.98$, $p = .005$); as well as in the sequential algebra SbS2 condition between $P(C_1 \mid E_i)$ and $P(C_1 \mid E_i, E_j)$ ($t(45) = -5.57$, $p < .001$) and between $P(C_2 \mid E_i)$ and $P(C_2 \mid E_i, E_j)$ ($t(45) = -6.13$, $p < .001$).

No significant differences in the sequential SbS1 condition.

Further analyses showed that none of the $P(C_1 \mid E_i, E_j)$ and $P(C_2 \mid E_i, E_j)$ are significantly different across the levels of the evidential learning condition whereas some $P(C_2 \mid E_i)$ are: in the full algebra condition $P(C_2 \mid E_i)$ in SbS2 and SbS1 are statistically different, $t(89) = -2.09$, $p = .04$, as well as $P(C_2 \mid E_i)$ in the sequential algebra condition SbS2 and SbS1 $t(88.5) = -2.51$, $p = .014$, with those in SbS1 having higher means. Combining these results from those above regarding participants estimates within each group suggests that (i) people are sensitive to the different orders the pieces of evidence of different diagonsticity were presented and (ii) that their estimates go against both the full CBN and the 'split' CBNs (qualitative) predictions since the differences $P(C_1 \mid E_i, E_j) - P(C_1 \mid E_i)$ and $P(C_2 \mid E_i, E_j) - P(C_2 \mid E_i)$ are larger in SbS2 condition than in SbS1 condition whereas the full CBN and the 'split' CBN predict exactly the opposite (see Figure 2.25).

A closer look at the data distributions in Figure 2.25 reveals the driving force of the results; namely, the participants' responses are highly clustered. Figure 2.26 shows that there are 5 significant clustering points. '.20', '.60', and '.70' seem to correspond to the probability values one finds in the CPTs for the effects. One clustering point ('.15') corresponds to the priors of the causes. The largest clustering point seems to be around the '.50' mark, which captured 34.1% of all answers to the test questions. Figure 2.27 shows the frequency of responses around ($\pm$.02) the five clustering points for each group. The data captured by the five clustering points amounted to 68.4% of all data.

Finally, to assess the fit of each model to the data, I calculated mean squared

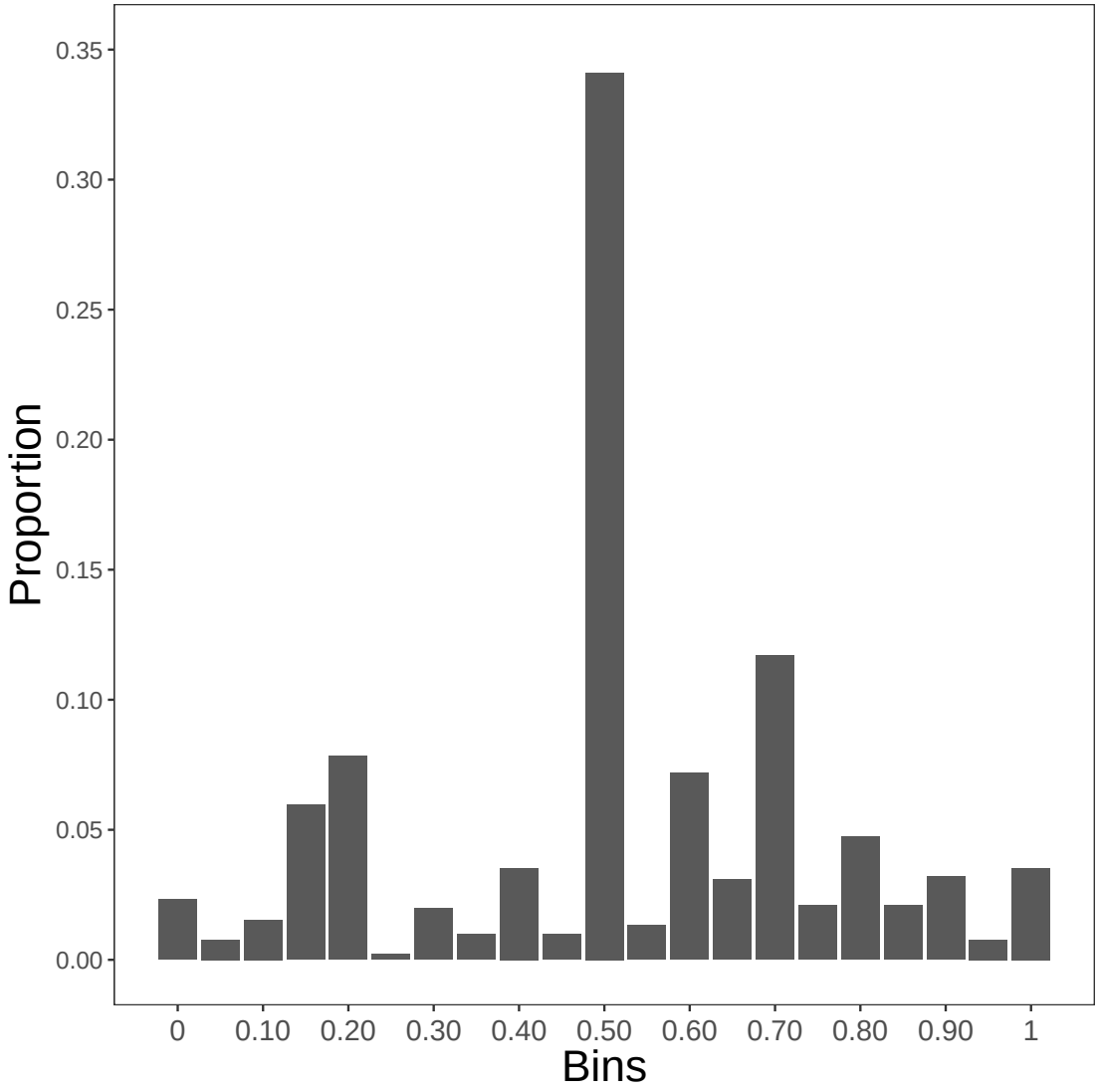


Figure 2.26: The distribution of all the test question estimates around the 21 clustering points (bins) ($\pm$ .02) in Experiment 4.

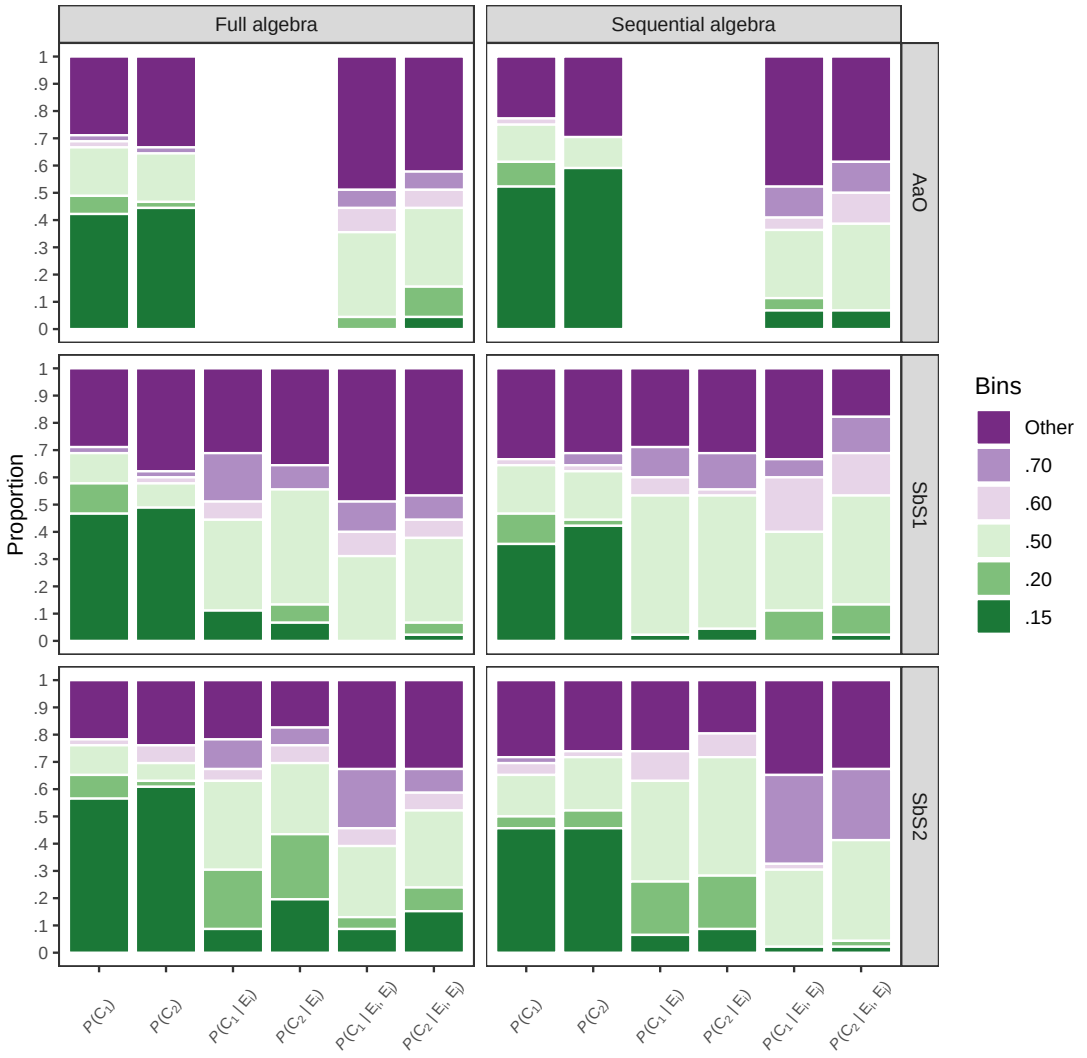


Figure 2.27: The frequency of participants' responses around the five focal points (bins) ($\pm$ .02) in Experiment 4.

errors (MSEs) for each model across the two algebra conditions.[29] Given the above-mentioned clustering around particularly the '.50' mark, I additionally calculated the MSEs for a simple model that included the correct priors (same as in both the full CBN and the 'split' CBN modeling), but has .50 as a response to all test questions. The results are presented in Table 2.16.

Table 2.16: MSEs for the full CBN, 'split' CBN, and '.50' model in the full and sequential algebra conditions

	Full algebra	Sequential algebra
Full CBN	621.18	536.94
'Split' CBN	778.93	701.97
'.50' model	573.73	496.65

The best fitting model of the three was the simple '.50' model, further confirming the clustering effect around the '.50' mark and the results of the linear mixed effect model. The full CBN model was a better fit than the 'split' CBN model of both the full algebra condition data and sequential algebra condition data. All three models fit better the sequential algebra condition data than the full algebra condition data suggesting a difference between the two conditions. However, the linear mixed effect model suggests that the difference is not statistically significant.

Diagnostic split Like in Experiments 1, 2, and 3, in this experiment we also have that $P(E_1 \mid \neg C_1, \neg C_2) = P(E_2 \mid \neg C_1, \neg C_2) = 0$; in other words, know-

[29] Note that the 'split' CBN does not have a unique prediction for AaO condition (see Figure 2.25). In calculating the MSE for that model I included the prediction that had the lower MSE.

ing that one of the effects has occurred makes the two independent causes exhaustive (but not mutually exclusive). Because of this, one would expect the diagnostic split reasoning to emerge in this experiment as well. Consequently, one would expect that after learning one of the effects has happened a number of participants would splits the probability space between the two causes such that $P(C_1 \mid E_i) + P(C_2 \mid E_i) = 1$. Similarly, one would expect that after learning the second effect has happened that some participant would estimate the probability of the causes such that $P(C_1 \mid E_i, E_j) + P(C_2 \mid E_i, E_j) = 1$. It is worth noting that even if participants provided estimates such that $P(C_1 \mid E_i) + P(C_2 \mid E_i) = 1$ and $P(C_1 \mid E_i, E_j) + P(C_2 \mid E_i, E_j) = 1$, it would not imply that they provided estimates such that $P(C_1 \mid E_i) = P(C_1 \mid E_i, E_j)$ and/or $P(C_2 \mid E_i) = P(C_2 \mid E_i, E_j)$. The participants could still be exhibiting diagnostic split reasoning even if their estimates for the two causes after learning the second effect differed to those they provided after learning only one effect has happened, as long as the probabilities of the two causes added up to 1.

To explore diagnostic split reasoning in this experiment, I calculated (i) the proportion of participants who provided their estimates in diagnostic reasoning after learning that one effect has happened such that $P(C_1 \mid E_i) + P(C_2 \mid E_i) = 1$ and (ii) the proportion of participant who provided their estimates after learning the second effect has happened such that $P(C_1 \mid E_i, E_j) + P(C_2 \mid E_i, E_j) = 1$. 37.9% of participants provided estimates ($\pm.02$) such that $P(C_1 \mid E_i) + P(C_2 \mid E_i) = 1$ and 31% of participants provided estimates ($\pm.02$) such that $P(C_1 \mid E_i, E_j) + P(C_2 \mid E_i, E_j) = 1$. Furthermore, 58% of the 37.9% of participants who provided estimates such that $P(C_1 \mid E_i) + P(C_2 \mid E_i) = 1$ went on to provide the probability estimates for the two causes that $P(C_1 \mid E_i, E_j) +$

$P(C_2 \mid E_i, E_j) = 1$. These results seem to suggest that diagnostic split reasoning was a significant driver of the probability estimate trends in data pertaining to the test question.

Propensity interpretation The propensity hypothesis also has a clear prediction in the context of this experiment. It predicts that some participants would stay at their prior probabilities and would not update the probabilities of the causes after leaning evidence occurred. More specifically, the hypothesis predicts that some participants would provide the estimates such that: $P(C_1) = P(C_1 \mid E_i) = P(C_1 \mid E_i, E_j)$ and $P(C_2) = P(C_2 \mid E_i) = P(C_2 \mid E_i, E_j)$. Only 7.4% of the participants provided estimates that matched these predictions of the propensity hypothesis. This suggests that only a small proportion of the participants in this experiment reasoned in agreement with the propensity hypothesis.

2.2.2.3 Discussion

In this experiment I found that people are sensitive to the order of presentation of the different pieces of evidence. However, although there was a trend in increasing the probabilities of the causes after finding out that the second piece of evidence obtained (in accordance with both the full and the 'split' CBN model), the (qualitative) predictions of both models regarding the amount of increase in each order go against the participants' mean estimates.

Further, I have explored people's diagnostic reasoning in the context of learning new variables. I found that people update almost identically when they are presented with the full algebra and when the algebra is expanded sequentially. In principle, this lack of difference could mean either that people are

very good at this expansion, or that they inappropriately treat the full model in a sequential, local fashion. The MSE analysis showed that the full CBN model is a better fit than the 'split' CBN model across board supporting the latter option. However, the significant clustering in our data and the fact that the '.50' model fitted the data better then either the full or the 'split' CBN model suggest that different participants employed different strategies in answering our test questions. Some of these seem indicative of well-established errors in human causal/probabilistic reasoning such as 'the inversion fallacy' where people confuse $P(\mathrm{A} \mid \neg \mathrm{B})$ with $P(\mathrm{B} \mid \mathrm{A})$ (Nance & Morris, 2002). Other strategies that participants seemed to have employed are discussed and experimentally tested in the first part of this chapter.

The propensity interpretation seemed to have played a small role in driving participants' probability estimates as only around 7% of them have provided estimates in accordance with the predictions of the propensity hypothesis. As I have speculated in the first half of this chapter, this could have well been because (i) the set-up in Experiment 4 was not deterministic, (ii) the cover story was not mechanistic in a sense that it did not include a mechanism of how causes bring about the effects, and (iii) the priors of causes were not established in any objective manner (for example, the prior probability of rain in this experiment was established by the protagonist in the cover story remembering a forecast that said that there was a 15% chance of rain overnight; in contrast, in Experiments 1 and 2 the prior probability of a coin to land Heads was established by a coin tossing mechanism that always tossed the coin with a particular probability for Heads). Given all these features of the cover story from Experiment 4, it then seems plausible that not many participants would

reason in accordance with the propensity hypothesis. Thus, (i), (ii), and (iii) provide useful guidelines for when very few people reason in line with the propensity interpretation.

The diagnostic split reasoning hypothesis has, however, accounted for a much larger proportion of participants' estimates. After learning that one of the effects has taken place, around 38% of the participants' estimates for the two causes could be accounted by the diagnostic split hypothesis and after additionally learning that the second effect has happened, around 31% of participants reasoned in line with the diagnostic split hypothesis. Although the above analysis seems to suggest that a significant number of participants engaged in the diagnostic split reasoning, this result would need to be further qualified. In Experiment 3, I explored whether participants who answered with .5 in diagnostic reasoning did so to express that they do not know what the answer is or their lack of confidence. The results from that experiment suggested that participants, indeed, did engage in the diagnostic split reasoning. Unlike in Experiment 3, in this experiment there was a very low percentage of the participants who correctly answered all the comprehension questions (that included both the questions about the priors and the CPTs). As the majority of the participants who seemed to have engaged in the diagnostic split reasoning have provided .5 as their estimate for the test questions, it is, thus, plausible that at least some of these participants wanted to communicate that they were not sure about their answer or that they did not know the answer.

The fact that a small proportion of participants answered all comprehension questions correctly, that a much greater proportion of them provided .5 as their estimate to test question, and that in general the participants' estimates

seemed more noisy than of those in the first three experiments all seem to suggest that participants struggled with the task in this experiment significantly more than in the first three experiments of this chapter. This could be for a number of reasons. For example, using a non-deterministic and/or a larger number of probability estimates that needed to be communicated to the participants could have resulted in a more noisy estimates and/or more .5 estimates that served as a stand-in for 'I don't know' or 'I'm not sure'. It could also be that the 4-node network is already too complex for the participants to provide meaningful estimates (this was suggested by the findings in Liefgreen et al., 2018). In the next experiment I explored one of these possibilities.

2.2.3 Experiment 5

The large proportion of participants who incorrectly answered the comprehension questions related to the priors and the CPTs in Experiment 4 may suggest that participants were overwhelmed with the quantitative information (in total 10 numerical values, 2 prior probabilities and 8 probability values for the CPTs). This may have resulted in a large proportion of participants simply repeating the quantitative information or replying with .50 (to communicate 'I don't know') to the test questions, thus obscuring any potential effects of extending the algebra on people's diagnostic reasoning estimates.

The goal of this experiment was to explore the effects of different orders of evidence presentation and sequential algebra expansion on diagnostic reasoning in situations where the amount of the quantitative information communicated to the participants was reduced and compare the empirical findings with the two predictions of both the full model and the split model.

This reduction was achieved in two ways: (i) participants were communicated only the strength of the causal relations between the effects and causes, more specifically they were communicated only the likelihood $P(E_j \mid C_i)$ for $i, j \in \{1, 2\}$;[30] and (ii) the strength of the causal relation was communicated in a verbal/qualitative manner rather than in a numerical/quantitative way.

Communicating only the strength of the causal relations reduced the number of parameters from 8 CPT values in Experiment 4 to four in this experiment. In total, in this experiment 6 parameters were communicated to the participants (2 priors and 4 probabilities regarding the strength of the causal relations). Reducing the number of parameters, however, comes at a cost. Using only the priors and $P(E_j \mid C_i)$ one cannot calculate the exact posterior probabilities $P(C_i \mid E_j)$ and $P(C_i \mid E_1, E_2)$ for either model. This is because one is missing the joint probability $P(E_1, E_2)$ in the case of the full model (E_1 and E_2 are independent in the split model so there $P(E_1, E_2) = P(E_1)P(E_2)$) and the probability of each effect given that either cause did not occur (i.e. $P(E_j \mid \sim C_i)$ for $i, j \in \{1, 2\}$) both for the split and the full model. Despite this, one can express the exact ratio of the posterior odds for the split model using only the likelihoods (see Appendix A.6). This ratio is:

$$\frac{P(C_i \mid E_i)}{P(C_i^* \mid E_j)} : \frac{P(C_j \mid E_i)}{P(C_j^* \mid E_j)} = \frac{P(E_j \mid C_j)}{P(E_j \mid C_i)} \tag{2.6}$$

The ratios in Equation 2.6 are then used to test the predictions of the split model. Unfortunately, the same cannot be done for the full model: in addi-

[30] Measuring causal strength as the likelihood $P(E_j \mid C_i)$ is not the only way to understand causal strength. Many other measures of causal strength have been proposed (for an overview see Fitelson & Hitchcock, 2011). Which of these measures is the most appropriate one is still quite a debated issue (see, for example, Chapter 6 of Sprenger & Hartmann, 2019).

tion to the likelihoods one would also need the joint probability over E_1 and E_2 (see Appendix A.6). Nonetheless, one can estimate where the corresponding ratio for the full model would be when compared to the split model ratios. This is done by incorporating the following information: (i) the posteriors of the causes after learning the first piece of evidence are the same for both models; (ii) the posteriors after learning the second piece of evidence are higher in the split model for both causes (for an illustration of (i) and (ii) see the predictions of the two models in Figure 2.25); (iii) the prior of the causes were unequal in this experiment (see below). Together, (i), (ii), and (iii) imply that the ratio of the posteriors for the full model is more likely to be in between the two ratios for the split model from Equation 2.6: i.e. between the split model ratio for when modeling situations where we first learn E_1 and then E_2 and the split model ratio for when first learn E_2 and then E_1.

The second modification of the experimental design in this experiment compared to Experiment 4 that aimed at reducing the amount of the quantitative information related to the manner the parameters were presented to the participants. Namely, all probabilistic information (the priors and the strengths of the causal relation, i.e. the likelihoods) were communicated to the participant through qualitative/verbal expression such as 'Rain almost always causes the lawn to be wet' or 'Sleep deprivation often causes magnesium deficiency'. Using only verbal expression to communicate the parameters further reduces the amount of the quantitative information.

We often communicate probabilities through verbal expressions Several empirical studies have investigated how verbal expression are interpreted (e.g. Harris & Corner, 2011;Weber & Hilton, 1990; for a recent review of the

literature on verbal probability expressions see e.g. P. J. Collins & Hahn, 2018). However, most of the studies on causal-probabilistic reasoning provided participants with the quantitative information, not least because the models that are used for comparison, such as CBNs, require numerical input. An exception is a study by Meder and Mayrhofer (2017a). They investigated people's sequential diagnostic reasoning with verbal information and compared it to the sequential diagnostic reasoning of those who were provided with the quantitative/numerical information. Specifically, the authors used a mapping of verbal expression to the numerical ones from Bocklisch, Bocklisch, and Krems (2012) and compared the sequential diagnostic reasoning of a group of participants who were communicated the priors and the likelihoods using verbal expressions (e.g. 'X is frequently the cause of the symptoms') to the sequential diagnostic reasoning of a group of participants who were communicated the same parameters using the mapped numerical expression (e.g. 'X is the cause of the symptoms in 67% of all cases'). The findings from Meder and Mayrhofer (2017a) suggested that there was a high consistency between the two groups and that both groups' diagnostic judgments were quite accurately tracking the predictions of a related Bayesian network model. In this experiment I relied on these findings. I communicated to the participant the 6 parameters in a verbal expression form and used the mapped numerical estimates from Bocklisch et al. (2012) to compute the ratio of odds for the split model.

Lastly, in this experiment I used low unequal priors. This was done in hope of reducing the number of .50 responses if they are a product of diagnostic split reasoning (in Experiment 2 I found that when reasoning with unequal priors people who engage in diagnostic split reasoning do not always provide .50 as

a response, but rather follow the ratio of the priors). Furthermore, the low unequal priors contribute to further narrowing down the prediction of the full model regarding the ratio of odds as mentioned above.

2.2.3.1 Overview

In this experiment, like in Experiment 4, I manipulated the order in which the evidence was presented and whether the full algebra was available to the participants from the start or the algebra was introduce sequentially. However, unlike in Experiment 4, in this experiment the all-at-once (AaO) condition was excluded. This was because (i) Experiment 4 suggested none of the $P(C_1 \mid C_i, E_j)$ and $P(C_2 \mid C_i, E_j)$ were significantly different across the evidential learning conditions, including AaO and (ii) given the parameters communicated to the participants, the two models had no specific predictions regarding this condition.

Further, two cover stories were used in this experiment: a modified version of the one from Experiment 4 to accommodate unequal priors and the verbal probability expression and a new one. The verbal probability expression used in this experiment (with their mapped numerical counterpart) were the following. For priors: $P(C_1)$ = 'unlikely' (.22) and $P(C_2)$ = 'very unlikely' (.12) (mapping from Wintle, Fraser, Wills, Nicholson, & Fidler, 2019). For the likelihoods I used the following expressions: $P(E_1 \mid C_1)$ = 'almost always' (.88), $P(E_1 \mid C_2)$ = 'often' (.7), $P(E_2 \mid C_1)$ = 'almost never' (.08), $P(E_2 \mid C_2)$ = 'sometimes' (.33) (mapping from Bocklisch et al., 2012).

Given these likelihoods, if participants reasoned in accordance with the split model we would expect that the ratio of their posteriors (from Equation 2.6)

would be around 4.1 when they are first told that E_1 obtains and then that E_2 obtains; when they are first told that E_2 obtains and subsequently that E_1 obtains this ratio would be around 1.3. For the full model, we would expect that the corresponding ratio for the full model will most likely be somewhere between these two values.

2.2.3.2 Methods

Participants and Design A total of 120 participants ($N_{\text{MALE}} = 48$, $M_{\text{AGE}} = 33.3$ years) were recruited from Prolific Academic (www.prolific.ac). All participants were native English speakers who gave informed consent and were paid £1.25 for partaking in the present study, which took on average 8.9 minutes to complete. Participants were assigned to one of the 2 (algebra: full or sequential) $\times$ 2 (evidence learning: step-by-step from E_1 to E_2 (SbS1), or step-by-step from E_2 to E_1 (SbS2)) = 4 between-participants groups (one group with 28 participants, 2 groups with 30 participants, and one group with 31 participants).

Materials In this experiment there were two cover stories. The first cover story ($n = 60$) was exactly the same as the one used in Experiment 4, except that all the information regarding the parameters was communicated to the participants through verbal expressions. The second cover story ($n = 59$) had sleep deprivation (C_1) and skipping magnesium rich foods (C_2) (two binary and independent variables) as potential causes of magnesium deficiency (E_1) and obesity (E_2). Unlike in Experiment 4, in this experiment there were no comprehension questions related to the CPTs of the effects. Thus, all participants were asked 6 questions: 2 about the priors of the causes and 4 test questions

relating to $P(C_1 \mid E_i)$, $P(C_2 \mid E_i)$, $P(C_1 \mid E_i, E_j)$, and $P(C_2 \mid E_i, E_j)$). For the full materials used in Experiment 5 see Appendix A.7.

Procedure The procedure was exactly like that of Experiment 4 for step-by-step levels of the evidence learning condition (i.e. excluding the all-at-once level) and the two levels of the algebra condition. The only difference was that participants were not asked any comprehension questions as these were not included in this experiment.

2.2.3.3 Results

Participants responses to the priors and the four test questions are in Figure 2.28.

Priors There was no significant difference between $P(C_1)$ and $P(C_2)$ in any of the four groups: for the full algebra, SbS1 group $t(29) = 0.97$, $p = .19$; for the full algebra, SbS2 group $t(30) = 1.75$, $p = .09$; for the sequential algebra, SbS1 group $t(29) = 1.36$, $p = .18$; and for the sequential algebra, SbS2 group $t(27) = -0.41$, $p = .69$.

Test questions To test the effect of the algebra and the evidence learning conditions on participants' estimates on the test questions, a linear mixed effects model was built. Like in Experiment 4, the model had two fixed effects, Algebra and Evidence learning, with a random intercept for each participant (there was no random slope for participant since algebra and evidence learning conditions vary between participants). There was no main effect of Evidence learning (though the p-value was quite close to $\alpha = .05$) and no main effect of Alge-

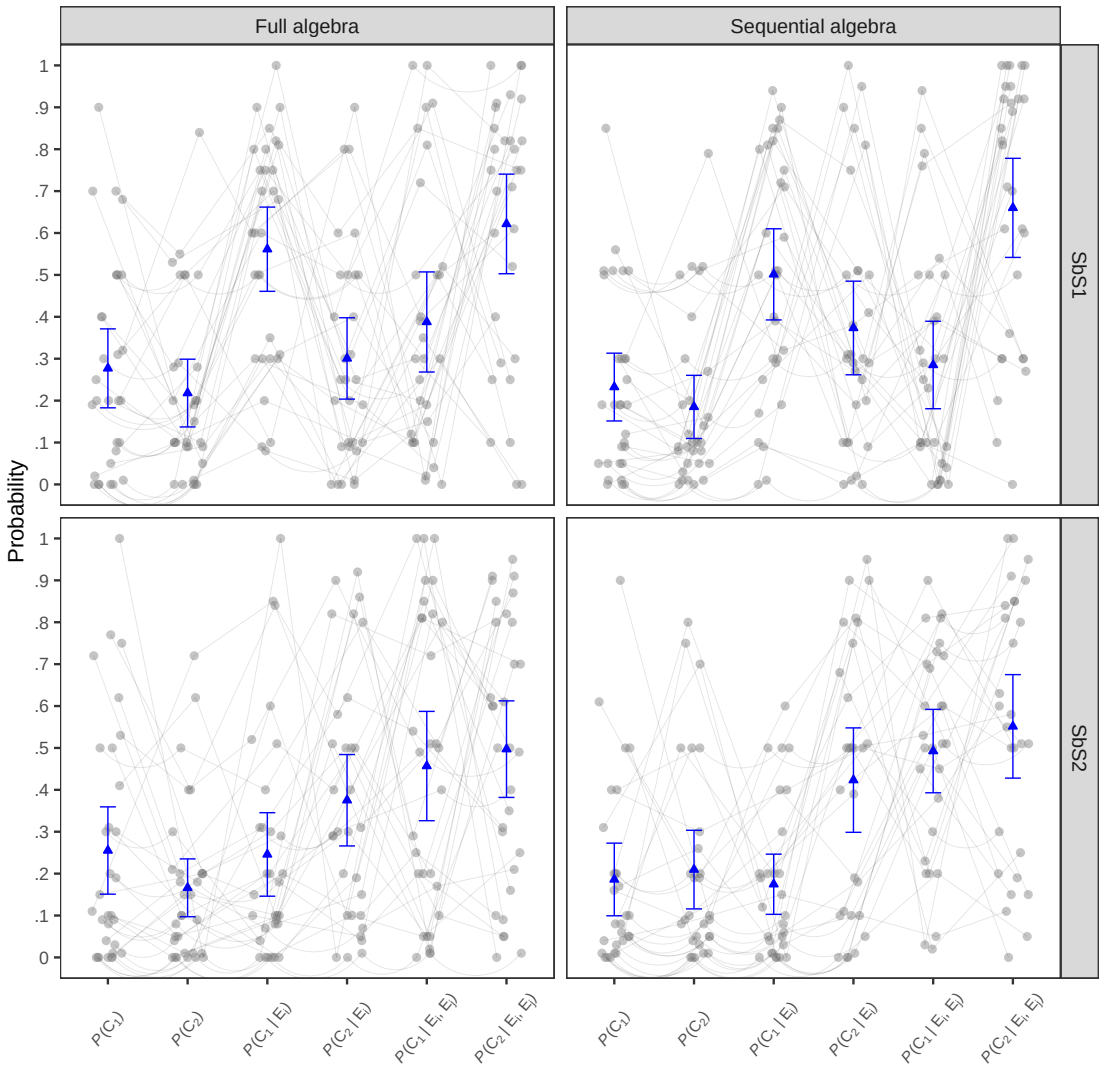


Figure 2.28: Responses of participants to priors and test questions in Experiment 5. Blue triangles are means and error bars are 95% confidence intervals. Gray lines between data points depict how participants changed their probability estimates from one questions to another, with curved lines indicating that a participant did not change (within $\pm$ .02) their probability estimate.

bra (see Table 2.17). I also found no interaction between Algebra or Evidence learning. Furthermore, likelihood ratio tests showed that including the predictors in the model does not improve model fit compared to just having an intercept as a predictor ($\chi^2(3) = 4.02$, $p = .26$). That is, the data grand mean fits the data no worse than the model which includes both predictors.

Table 2.17: Linear mixed effect model results for test questions in Experiment 5
A=Algebra; EL=Evidence learning

	Estimate	95% CI	t-value	p
A	0.02	[-0.58, 0.62]	0.06	.95
EL	-3	[-5.98, 0.05]	-1.9	.058
A$\times$ EL	0.14	[-0.45, 0.75]	0.48	.63

A finer grained analyses on the data within each group showed the following results. Further, a significant difference was found both between $P(\mathrm{C}_1 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_1 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(29) = 3.8$, $p < .001$) and $P(\mathrm{C}_2 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_2 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(29) = -6.03$, $p < .001$) in the full algebra SbS1 condition; in the full algebra SbS2 condition there was a significant difference both between $P(\mathrm{C}_1 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_1 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(30) = -3.6$, $p = .001$) and $P(\mathrm{C}_2 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_2 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(30) = -2.5$, $p = .018$); in the sequential algebra SbS1 condition there was a significant between $P(\mathrm{C}_1 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_1 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(29) = -4.2$, $p < .001$) and between $P(\mathrm{C}_2 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_2 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(29) = -5.13$, $p < .001$); in the sequential algebra SbS2 condition there was a significant between $P(\mathrm{C}_1 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_1 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(27) = -6.3$, $p < .001$) and between $P(\mathrm{C}_2 \mid \mathrm{E}_i)$ and $P(\mathrm{C}_2 \mid \mathrm{E}_i, \mathrm{E}_j)$ ($t(27) = -3$, $p = .006$). These results imply that learning a second piece of evidence changed participants probability estimates compared to when they knew

about only one piece of evidence.

Further analyses showed that none of the $P(C_2 \mid E_i, E_j)$ were significantly different across the levels of evidential learning and only in the sequential algebra was there a difference in $P(C_1 \mid E_i, E_j)$ between the two levels of the evidential learning condition $t(56) = -2.9$, $p = .005$. None of the $P(C_2 \mid E_i)$ were significantly different across the levels of evidential learning, whereas all $P(C_2 \mid E_i)$ were: in the full algebra condition $P(C_1 \mid E_i)$ for SbS2 and SbS1 were statistically different, $t(59) = 4.6$, $p < .001$, and in the sequential algebra condition $P(C_1 \mid E_i)$ for SbS2 and SbS1 were statistically different $t(50) = 5.1$, $p < .001$. Combining these results from those above regarding participants estimates within each group suggests that people are sensitive to the different orders the pieces of evidence of different diagonsticity were presented, but that there was no effect of the algebra condition. These results very closely resemble those from Experiment 4.

The prediction of the split model regarding the ratio of the priors from Equation 2.6 was that in SbS1 evidential learning condition this ratio would be around 4.1 and in SbS2 this ratio would be around 1.3, and the ratio for the full model would most likely be in between these two values. I have calculated the odds ratio for each participants' (removing those whose ratios required division by 0) and compared it to split model predictions. For the group reasoning in the full algebra SbS1 conditions there was no significant difference between split model prediction and the participants' derived ratios (the mean ratio was 14.6), $t(24) = 1.85$, $p = .077$; similarly for the group reasoning in the sequential algebra SbS1 conditions (the mean ratio was 12.4), $t(24) = 1.94$, $p = .063$; for the group reasoning in the full SbS2 conditions there was no

significant difference between the split model prediction and the participants' derived ratios (the mean ratio was 1.9), $t(28) = 1.04$, $p = .3$; lastly, for the group reasoning in the sequential SbS2 condition this difference was significant (the mean ratio was 0.66), $t(24) = -6.75$, $p < .001$. These results seem to partially support the split model's predictions. However, given that (i) the means for the SbS1 groups were very close to being significantly higher, (ii) the mean for the sequential SbS2 group was significantly different from the prediction, and (iii) there was no significant difference between the two algebra conditions, the support for this model is very limited. In addition, the results do not support the (range of) predictions from the full model.

Unlike in Experiment 4 where we saw a high clustering of participants' responses (particularly around the '.50' mark), in this experiment that is much less the case. Only 11.3% of all answers to the test questions were .50 ($\pm$.02) (see Figure 2.29). This is significantly lower than 34.1% which was in Experiment 4. All other clusters in this experiment capture around 10% of the data or less. This suggests that there was no significant clustering in Experiment 5 that could account for the large proportion of the participants' estimates and that the experimental design in this experiment was successful in reducing the clustering (especially around the '.50' mark).

Diagnostic split To explore how many participant provided estimates to the test questions in line with the diagnostic split reasoning, I calculated the proportion of participants whose estimates were such that $P(C_1 \mid E_i) + P(C_2 \mid E_i) = 1$ and $P(C_1 \mid E_i, E_j) + P(C_2 \mid E_i, E_j) = 1$. Only 16.8% of participants provided estimates such $P(C_1 \mid E_i) + P(C_2 \mid E_i) = 1$ and 12.6% of participants provided estimates such $P(C_1 \mid E_i, E_j) + P(C_2 \mid E_i, E_j) = 1$. These pro-

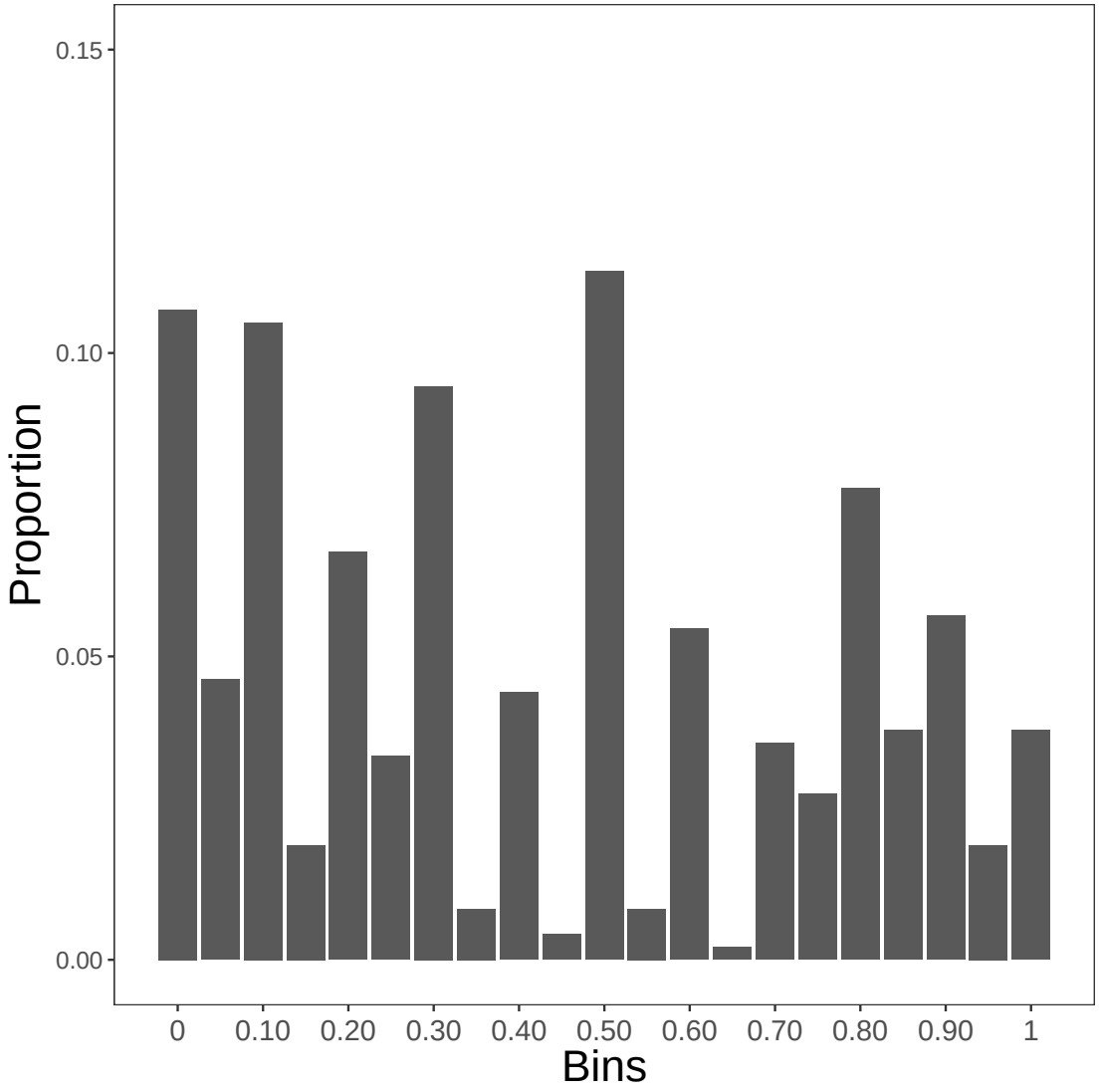


Figure 2.29: The distribution of all the test question estimates around the 21 clustering points (bins) ($\pm$ .02) in Experiment 5.

portions are significantly lower than in Experiment 4 where the corresponding proportions were 37.9% and 31%.

Propensity interpretation Only 4.2% of participants provided estimates such that: $P(C_1) = P(C_1 \mid E_i) = P(C_1 \mid E_i, E_j)$ and $P(C_2) = P(C_2 \mid E_i) = P(C_2 \mid E_i, E_j)$. This was lower than in Experiment 4 where this proportion was 7.4%. This results that only a small proportion of the participants in this experiment reasoned in agreement with the propensity hypothesis.

2.2.3.4 Discussion

Overall, the results from Experiment 5 closely resemble those from Experiment 4. There was no significant difference between the algebra conditions and there seems to have been an influence of the order in which the evidence was learnt. These findings persisted despite the new design where the probabilities were communicated through verbal expression and despite no significant clustering around the probability estimates (especially around the '.50' mark) compared to Experiment 4.

Although only the strengths of the causes (i.e. the likelihoods) were communicated to the participants, one could derive the ratio of the posterior odds of the split model and express it via as the ratio of the likelihoods, that was then used as a prediction of the split model. To calculate these likelihoods I used the mapped numerical estimates of the verbal probability expressions. A limited support for the split model and no support was found for the full model was found. However, there are further limitations to this support. Using only the data collected from Experiment 5 one cannot determine whether the participants (on average) accepted the mapped numerical estimates of the verbal

probability expressions. Furthermore, as studies have documented large individual differences in understanding, communicating, and using verbal probability expressions, it is then even more difficult to determine whether the participants accepted the mapped numerical estimates (see, for example, Wallsten & Budescu, 1995). The finding that participants' estimates seemed to have reflected different causal strengths in different orders of evidential learning suggests that participants were sensitive to these different strengths that were communicated via the verbal expressions. However, it is not clear to what extent they were sensitive to these different expressions, as, for instance, there was no significant difference between the two average estimates for the priors in any of the four groups.

The diagnostic split and the propensity hypothesis were also explored in this experiment. The results suggested that the proportion of people who reasoned in line with the diagnostic split hypothesis was significantly lower than in Experiment 4 and somewhat lower for the propensity hypothesis. Some of the potential reasons for this finding regarding the diagnostic split hypothesis are that (i) participants were not communicated/it was left of to interpretation whether the two causes became exhaustive after learning the effects has happened (as was the case in the first four experiments), which could then leave open the possibility for unobserved causes that could have also resulted in the effect occurring, and/or (ii) the verbal probability expressions do not trigger diagnostic split reasoning as much as the numerical probability estimates. The latter explanation could also account for the somewhat lower proportion of the participants who reasoned in line with the propensity interpretation in Experiment 5 compared to Experiment 4. These should be explored in future research.

2.2.4 General discussion

The general goal of the part of the chapter was twofold. First, I sought to explore new avenues in sequential diagnostic reasoning by investigating people's causal judgements with multiple independent causes and multiple pieces of evidence of different diagnosticity. To this effect I found that people are sensitive to the different diagnosticity of the evidence. This was the case when the probabilistic information regarding the diagnosticity was communicated both through the numerical probability estimates and the verbal probability expressions. However, I have found that people do not follow the predictions of the normative (full) CBN model and sometimes provided estimates that went against the (even qualitative) prediction of the normative model. This suggests that although people were sensitive to the different diagnosticities of the evidence, it is perhaps not these diagnosticities that led them to provide specific estimates for the two causes. In contrast, Meder and Mayrhofer (2017a) found that both the numerical probability estimates and the verbal probability expression were effective in communicating the different diagnosticities of the effects in the case of the mutually exclusive causes.

The second goal was to introduce the issue of novel variables in sequential reasoning and the practical as well as modeling challenges it presents. In particular, I have explored two modeling strategies for extending algebra in the context of adding an effect to a common-effect 3-node causal structure: a split CBN model and a full CBN model. I have shown that predictions of the two models differ under most of the circumstances. The results from Experiments 4 and 5 have, however, suggested that (i) there was no difference between the groups who were presented the full algebra from the beginning and the groups

who were presented the algebra in a sequential way and (ii) that there was little support for either the split or the full CBN model.

One potential explanation of these results is that there actually is no difference between reasoning with full algebra or learning the algebra in a sequential manner. The results from Experiment 4, however, showed a high degree of clustering around specific focal points (in particular '.50' mark), which seemed to have driven participants' estimates. By reducing the amount of the quantitative information and using verbal probability expression the clustering has significantly reduced in Experiment 5. However, even in Experiment 5 participants seemed to have strongly relied on the likelihoods to provide estimates to the test questions, with the clustering being less pronounced because of the vagueness of the verbal probability expressions. This suggests at least two possible explanations: (i) the experimental design used in Experiment 4 and 5 is not appropriate to explore diagnostic reasoning in more complex (2 independent causes and 2 effects) causal situations or (ii) the 4-node causal situations with 2 independent causes and 2 effects are already too complex for people to process and reason with. The latter explanation is not without merit as the recent research suggests that, when presented with more complex situations, people preform poorly compared to the normative CBN models (see e.g. Cruz et al., 2020; Liefgreen et al., 2018). This does not imply that all 4-node causal structures are too complex for people to reason with. For instance, research on diamond structures where there is a one common cause of two further intermediate causes which in turn cause one common effect suggests that people in general demonstrate the basic normative patterns when reasoning with these kinds of structures (Meder, Hagmayer, & Waldmann, 2008, 2009). What poten-

tially sets apart the structure used in Experiment 4 and 5 from the diamond structure is that the structure in Experiment 4 and 5 includes two explaining away patterns (compared to only one in the diamond structure) that people find in general difficult process as discussed in the first part of the chapter and that it has one extra parameter. All this could result in the complexity that goes beyond people's reasoning abilities.

The two hypotheses from the first part of the chapter, the propensity interpretation and the diagnostic split, were explored in this part of the chapter as well. The results suggested that the diagnostic split hypothesis accounted for a large proportion of the responses in Experiment 4, although some of these responses could have been due to participants' providing .50 as their estimate to communicate that they do not know the answer. In Experiment 5, the diagnostic split hypothesis account for a significantly smaller proportion of the estimates, suggesting that (i) the diagnostic split reasoning is less prevalent when the parameters are communicated as verbal probability expression and/or (ii) the fact it was left open to the participants to interpret whether there are any unobserved causes that could have also produced the effects might have lead the participant away from the diagnostic split reasoning. The propensity interpretation accounted for a significantly smaller proportion of participants estimates in Experiments 4 and 5 compared to Experiments 1-3. This could be because, as speculated in the first part, in Experiments 4 and 5 (i) the set-up was not deterministic, (ii) the cover stories were not mechanistic, and (iii) the priors were not established in an objective way. The even lower proportion of the estimates that aligned with the propensity interpretation in Experiment 5 compared to Experiment 4 could be because of the parameters being communi-

cated through verbal expressions. The vagueness of the verbal expression than could have even more contributed to, for instance, the priors not been clearly and objectively established. Further research should explore these avenues.

2.3 Conclusions

In this chapter I have explored how explanation when thought of as a product and explanation of a specific outcome in an intrapersonal context can affect our beliefs in causal reasoning. Specifically, I explored a particular causal reasoning pattern explaining away. I have found that when engaging in basic 3-node explaining away situations people are significantly deviating from the CBN normative prediction, but that they are doing that in a very specific ways that could be captured by the two hypotheses: the diagnostic split and the propensity interpretation.

I have also explored sequential diagnostic reasoning in explaining away situations with multiple pieces of evidence. Here, I also explored the potential effects of sequentially extending algebra. This was a novel exploration from both a modeling and an empirical perspective. I have found that people are sensitive to different diagnosticities of evidence, but no effects of the algebra extension were found. As the research on extending algebra is still in its infancy regarding both the modeling strategies and the empirical exploration, further research is warranted.

If one were to take a step back from all the details of the experiments in this chapter and consider the notion of explanation employed in this chapter they would find that this notion is quite limited. The explanation to consists of one node in a CBN that is 'responsible' for a change in the probability of another

node. This, however, is not the only way to understand explanations. In the next chapter I discuss the different notions of explanation and their relations to, in particular, the AI literature.

3

Explaining the argument: The case of causal Bayesian networks

In the previous chapter I explored the relationship between argument/belief change and explanations of specific outcomes. However, we often ask for or are invited to provide explanations of not just the presence or existence of specific outcomes, but also how these outcomes came about. For example, a doctor

could point to a specific disease as an explanation of why a patient has particular symptoms. The patient, however, could ask a doctor why do they think that the symptoms could be accounted for by the disease. In this case, a doctor (or more broadly a scientific community) would need to provide an explanation of an inference or a reasoning process(es) that led them to think that a specific disease is a cause of particular symptoms. Similarly, in court it is often not sufficient just to say that a particular person is guilty of a crime. Often, a prosecutor is required to present and explain a case or an argument or an inference process regarding how the evidence is incriminating the defendant.

In more recent times, explaining reasoning and decision-making processes of the artificial intelligence (AI) systems has emerged as a prominent issue. For example, an Amazon recommender system, an AI system that provides automatic recommendations, could recommend to us a particular book. However, as many of us have experienced, it is often unclear why a particular book is recommended. Further explanation of why that particular book was recommended and of what decision-making process led the AI system to this recommendation is, either explicitly or implicitly, required by the human user if that recommendation is to have an impact on whether or not the user will ultimately buy the book. The field called 'the explainable AI' (XAI) is devoted to addressing this challenge of explaining the decisions and inference processes of AI systems. Furthermore, recent years have seen a groundswell of interest in machine-generated explanation for AI systems, where (one) AI system is generating explanations of (another) AI system's inference processes (Doshi-Velez & Kim, 2017; Gunning & Aha, 2019; Montavon, Samek, & Müller, 2018; Rieger, Chormai, Montavon, Hansen, & Müller, 2018; Samek, Wiegand,

& Müller, 2017). Multiple factors exert pressure for supplementing AI systems with explanations of their outputs. Explanations provide transparency for what are often black-box procedures. Hence transparency is viewed as critical for fostering the acceptance of AI systems in real-world practice (Bansal, Farhadi, & Parikh, 2014; Chen et al., 2014; Fallon & Blaha, 2018; B. Hayes & Shah, 2017; Mercado et al., 2016; Wachter, Mittelstadt, & Russell, 2017), last but not least, because transparency might be a necessary ingredient for dealing with legal liability (Felzmann, Villaronga, Lutz, & Tamò-Larrieux, 2019; Doshi-Velez et al., 2017; Goodman & Flaxman, 2016; Wachter, Mittelstadt, & Floridi, 2017). At the same time, decades of research in AI make plausible the claim that AI systems genuinely able to navigate real-world challenges are likely to involve joint human-system decision making, at least for the foreseeable future. This however, requires AI systems to communicate outputs in such a way as to allow humans to make informed decisions. In other words, it would require human-friendly explanations of AI systems' reasoning and decision-making processes that led the AI systems to produce a particular outcome or recommend a particular action.

In this chapter, I will explore the explanations of inference processes or arguments. The kinds of explanations that I am going to be considering are still made in an intrapersonal context and are products of explanation processes. Figure 3.1 illustrates where these explanations are located in the 3-dimensional explanation cube.

The focus of this chapter will be explaining inference processes in causal Bayesian networks (CBNs). There are two main reasons to concentrate on explaining inference in CBNs. Firstly, in Chapters 1 and 2 we have seen that

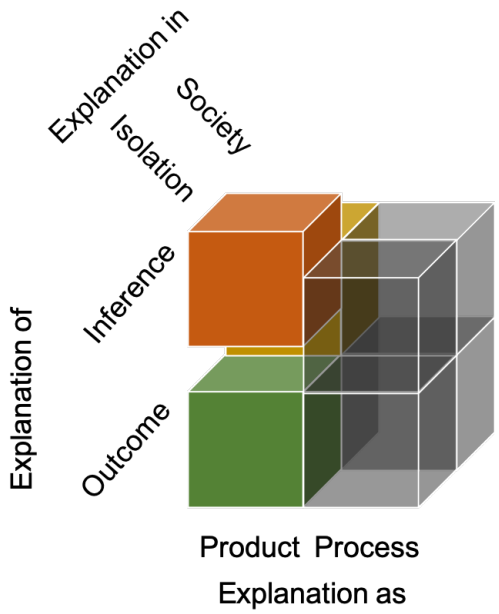


Figure 3.1: The three dimensions of explanation.

many argument schemes can be modeled using CBNs. Explaining inferences in concrete CBNs would thus equate to explaining arguments themselves, which would further inform the discussion on the relationship between arguments and explanations. Secondly, CBNs are an AI technique that has been viewed as significantly more interpretable and transparent than deep neural networks (Gunning & Aha, 2019), while still possessing a notable predictive power and being applied to various contexts ranging from defence and military (Falzon, 2006; Laskey & Mahoney, 1997; Lippmann et al., 2006) and cyber security (Chockalingam, Pieters, Teixeira, & van Gelder, 2017; Xie, Li, Ou, Liu, & Levy, 2010), over medicine (Agrahari et al., 2018; Wiegerinck, Burgers, & Kappen, 2013), and law and forensics (Lagnado, Fenton, & Neil, 2013; Fenton, Neil, & Lagnado, 2013), to agriculture (Drury, Valverde-Rebaza, Moura, & de An-

drade Lopes, 2017) as well as psychology and philosophy as seen in previous chapters. As such, CBNs are a promising meeting point connecting the research on machine-generated explanation in AI and the research on human understanding of explanation in psychology and philosophy. Furthermore, given the increasing popularity of CBNs within AI (Friedman, Geiger, & Goldszmidt, 1997; Pernkopf & Bilmes, 2005; Roos, Wettig, Grünwald, Myllymäki, & Tirri, 2005; Ng & Jordan, 2002), including their relation to deep neural networks (Choi, Wang, & Darwiche, 2019; Rohekar, Nisimov, Gurwicz, Koren, & Novik, 2018; Wang & Yeung, 2016) and efforts to explain deep neural networks via CBNs (Harradon, Druce, & Ruttenberg, 2018), explaining inferences in CBNs would not just inform work in XAI, but also more generally in AI.

In order to explore what is explanatory in the context of CBNs (or any other context), one would need some theoretical background regarding what counts as an explanation and, more specifically, what counts as a 'good' explanation. In Chapter 1 we have seen that there are many measures of argument quality. Similarly, there are also many ways to explicate explanatory goodness. In this chapter I review some of these ways and apply them to the case of explaining inference in CBNs.

The chapter proceeds in two parts. In the first part I review different notions of explanation with a specific focus on CBNs and introduce key theoretical perspectives on what constitutes an explanation, and more specifically a 'good' explanation, from the philosophy literature. In the second part I compare these theoretical perspectives and the criteria they propose with a case study on explaining reasoning in CBNs.

3.1 Theoretical background[1]

In the previous chapter I have explored a particular reasoning schema called 'explaining away' where learning a state of a variable in CBN is sufficient to 'explain away' the evidence and thus reduce the probability of other possible causes. In that chapter I have only briefly mentioned what is meant by an explanation in the context of explaining away, without going into a broader discussion on explanation. This was because the main goal of that chapter was to explore human causal reasoning in explaining away and its extensions and to address specific modeling strategies. In this part of the chapter, however, I am going to do discuss different notions of explanation that are found in the literature and that also apply to CBNs. Taking a step back from looking at a specific reasoning schema, in this part of the chapter I will discuss how the notion of explanation in the case of explaining away (as well as in other reasoning patterns) relates to the notion of explanation of inference.

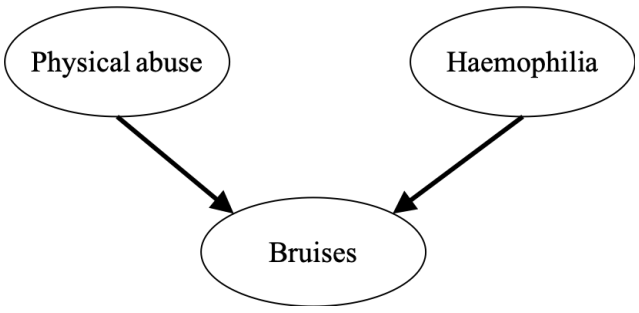


Figure 3.2: A CBN model of explaining away.

[1]This section is based on work from Tešić and Hahn (in press).

3.1.1 Explaining evidence

Defining an explanation has proven to be quite a difficult task (Lombrozo, 2012). Explanations have been understood as answers to how- and why-questions, as well as judgments about why an outcome occurred or hypotheses that include causes of what is being explained (also known as explanandum). These different ways of delineating an explanation point to different aspects of an explanation. Generally speaking, however, explanations can be understood as proposition that address a request for an explanation (Lombrozo, 2012). For instance, in the case of the inference of the best explanation (IBE) one is aiming to find the best explanation (hypothesis/proposition) of evidence (or data); once such explanation is found it then warrants a conclusion suggested by an explanation (Harman, 1965; Lipton, 2003). To illustrate, consider again an example where a doctor is presented with a child who is in pain (evidence/data). To decide an appropriate treatment, the doctor needs to to answer the following question: why is the child in pain? (request). The doctor finds that the best explanation of the child's pain is that the child has pulled a muscle (hypothesis/proposition) (the example is due to Okasha, 2000).

This notion of explanation maps well onto CBNs. In Chapter 2 we saw how, in explaining away situations, learning that one cause (hypothesis/proposition) has happened is sufficient to explain the effect (evidence, but also, as a consequence, can affect the probability of another (initially) independent cause. To use the example introduced in that chapter and graphically represented in Figure 3.2, learning that a child has haemophilia (hypothesis/proposition) is sufficient to explain the bruises on the body (evidence) and as a consequence would reduce the probability of physical abuse. Explain-

ing away reasoning is an example of intercausal reasoning, one of the three main types of reasoning in CBNs. As noted earlier in this thesis, the other two types of reasoning are predictive reasoning and diagnostic reasoning (often referred to as an abduction, see, e.g., Korb & Nicholson, 2010). Consider again the CBN in Figure 3.2. An example of predictive reasoning would be inferring a probability that the child has bruises given that it's suffering from haemophilia (i.e. inference from causes (or hypotheses, h) to effects (or evidence, e)); whereas diagnostic reasoning would be inferring a probability of physical abuse from learning that the child has bruises (i.e. inference from effects (e) to causes (h)). In all these three types of reasoning, an explanation is a variable (or a proposition) that can account for the presence/absence of another variables (evidence) in a CBN and that often has an impact on the probability of other variables in a CBN. This is a very simple view, but thinking of explanations as consisting of a variable that can account for the changes in states of other variables in a CBN does resemble similar ideas from psychology and philosophy regarding explanation more generally.

Other ways of explaining evidence in CBNs go beyond this simple view of explanation, often, however, building upon it. For example, diagnostic reasoning (abduction) is used to find the most probable explanations (causes) of observed evidence (effects), that is, to find the configuration h with the maximum $p(\mathrm{h} \mid \mathrm{e})$ (the approach is called 'maximum a posteriori' (MAP) and is due to Pearl, 1988). Similarly, Shimony's (1991) partial abduction approach first marginalizes out variables that are not part of explanations (x) and then searches for the most probable h: that is, find h with the maximum $\sum_x p(\mathrm{h}, \mathrm{x} \mid \mathrm{e})$. More recently, Yuan, Lim, and Lu (2011) introduced a method

they call 'Most Relevant Explanation' (MRE) which chooses the explanation that has the highest likelihood ratio compared to all other explanations: that is, find h with the maximum $p(\mathrm{e} \mid \mathrm{h})/p(\mathrm{e} \mid \overline{\mathrm{h}})$, where $\overline{\mathrm{h}}$ denotes all other alternative explanations to h. Nielsen, Pellet, and Elisseeff (2008) introduced a 'Causal Explanation Tree' (CET) method which uses the post-intervention distribution of variables (Pearl, 2009) in selecting explanations, which is in contrast to all previous methods since they use a non-interventional distribution of variables in a CBN. Drawing on their definition of causation, Halpern and Pearl (2005b) develop a definition of explanation to address the question of why certain evidence holds given users epistemic state. Their definition of explanation states that (i) a user should consider evidence to hold, (ii) an explanation (h) is a sufficient cause of evidence, (iii) h is minimal (i.e. it does not contain irrelevant or redundant elements), and (iv) h is not known at the beginning, but it is considered as a possibility. This is an improvement over other accounts. However, their account of causation, again, has as an output a set of variables in a CBN model which is deemed the causes of evidence in the model. Yap, Tan, and Pang (2008) employ the Markov blanket to determine which variables should feature in an explanation. A Markov blanket of a node A includes all nodes that are direct parents, children, or children's parents of that node. For example, all nodes within the dashed circle in Figure 3.3 constitute a Markov blanket of node A. A powerful property of the Markov blanket is that knowing the states of all the variables in a Markov blanket of A would uniquely determine the probability distribution of A: additionally learning the states of other variables outside the Markov blanket of A would not affect the probability distribution of A. Yap et al.'s 'Explaining CBN Inferences' procedure identifies Markov

nodes of evidence (i.e. nodes in a Markov blanket of the evidence node) and learns context specific independences in Markov nodes with a decision tree to exclude irrelevant nodes in an explanation of the evidence.

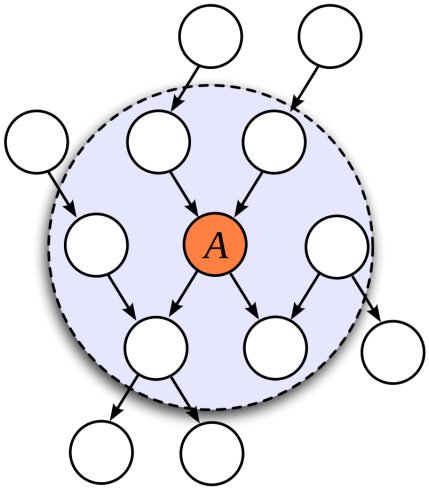


Figure 3.3: An example of a Markov blanket. All nodes within the dashed circle constitute a set of nodes that is a Markov blanket of node A. The illustration is publicly available via Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/File:Diagram_of_a_Markov_blanket.svg

Despite the differences among these methods, they all share at least one commonality: explanation of evidence is exhausted by a set of variables (hypotheses/propositions) in a CBN that these methods have pointed to. In other words, explanatory justification is provided in terms of a set of variables. This observation is not, however, restricted to explanations in CBNs. Image classification is a problem commonly addressed in deep neural networks literature. XAI approaches to image classification in deep learning are looking for pixels that contribute the most to an image being classified under a certain category (Lundberg & Lee, 2017). Although this is an important and commendable re-

search stream in XAI, it does not address how an AI system came to a conclusion (output) and it does not provide an explanation of the inference processes in an AI system: why is it that certain pixels contributed more than the others to an image being classified under a certain category? Why is it that a certain subset of pixels is sufficient for an AI system to classify an image? (Sun, Chockler, Huang, & Kroening, 2019).

Selecting a set of nodes (or pixels) that best justifies the evidence (output) according to a particular method or selecting a hypotheses that have the highest explanatory power is undoubtedly useful. However, in certain contexts (e.g. high-cost domains, see Herlocker, Konstan, & Riedl, 2000) this is arguably not enough to meet the demands of user transparency. In contrast to the notion of explanation as a justification of evidence is that of explanation of *reasoning processes* in CBNs, and expert systems in general (Lacave & Díez, 2002; Sørmo, Cassens, & Aamodt, 2005; Wick & Thompson, 1992). Here one is interested in how evidence propagates in a CBN rather than in selecting a set of variables that would account for evidence.

3.1.2 Explaining reasoning processes

Sometimes just finding single propositions or sets of propositions is not sufficient to respond to the explanation request. Often, we are required to explain our *reasoning processes* that led us to believe that a particular proposition is an adequate explanation. For example, in court proceedings an explanation regarding why certain a person A died is that the defendant killed them. But simply saying that the defendant killed A is usually not enough for a judge to convict the defendant of the crime. The prosecution is required to present a case

or make an argument that further elaborates on why they think the defendant killed A, *connecting* the pieces of evidence to the claim that the defendant killed A and explaining how the evidence bears on the defendant's guilt. Thus, in this example the original claim (the defendant killed A) is not sufficient to explain the evidence and we are in the search for further justification and clarification that goes beyond data and evidence to link that data to the claim/hypothesis. The explanation in this case would consist of the justification and clarification that connects the evidence and the hypothesis.

The view that explanations are links between claims/hypothesis and data/evidence is not novel. For instance, Toulmin in his book *The uses of argument* introduces his highly influential argument framework (Toulmin, 1958/2003). There, he differentiates between claims, data, and warrants. Claims are propositions we are trying to argue for using data. For instance, the fact (data) that Harry's hair is red provides direct support for the claim that Harry's hair is not black. However, the way data provides support for a claim is not always obvious. For example, it may not be clear to everyone that the fact (data) that a person is a Swede provides support for the claim that they are not Roman Catholic; sometimes a further warrant or explanation is needed, such as 'A Swede is most likely not a Roman Catholic' to connect the data to the claim. Thagard (1989) argues even more forcefully for the conception of explanations as links between data and claims (or hypotheses) and Antaki and Leudar (1992) view explanations as providing support for claims. More recently, Brem and Rips (2000) explored whether people are able to distinguish between explanation and evidence. They distinguish between the claim ('a proposition whose truth value we are attempting to establish'), evidence (data), and expla-

nation that can provide support for the claim by providing a (causal) bridge between data and the claim. For instance, one could argue that welfare recipients have difficulty getting off public aid (data/evidence) because they lack job skills (claim). However, it is not necessarily obvious how lack of skills could lead to the difficulty of getting off the public aid. To that end one could provide an explanation such as 'Job skills increase a person's chances of landing a well-paid job, which in turn supplies them with enough money to give up welfare checks'. This explanation further elucidates the relationship and an inference process between the claim and data and as such provides a bridge between them.

The conception of explanations as bridges or links between claims and data is not only found in the psychology and philosophy literature. At the beginning of this chapter I mentioned an example regarding an Amazon recommender system providing us with a book recommendation. When an AI system provides us with a recommendation (e.g. an Amazon recommender system recommending us a book to buy) based on our search history (data), to most of us users it is not clear how the AI system came to a particular recommendation. One of the challenges in computer science and, more specifically, in recommender systems research, is not just finding a recommendation for a human user, but also providing a machine-generated explanation of that recommendation (Zhang & Chen, 2020). The goal is to build AI systems that would generate explanation for the recommendations (e.g. the book has been recommended because it is in the same category as the books we previously bought), which would provide users with an insight into how the AI system came to its recommendation so that users are more likely to trust the recommendation and

follow up on it. Here, from a user perspective explanations can also viewed as links between data and a recommendation: an explanation provides an insight into how the AI system came to a particular recommendation given the data, that is it provides an insight into the reasoning processes of an AI system.

Explaining reasoning processes in CBNs has also been a research focus amongst expert systems researchers for some time (see Lacave & Díez, 2002 for an overview). I describe explaining reasoning processes in CBNs through a recent attempt in the context of the Bayesian Argumentation via Delphi (BARD) project.

The BARD project (Cruz et al., 2020; Dewitt, Lagnado, & Fenton, 2018; Liefgreen et al., 2018; Nicholson et al., 2020; Phillips, Hahn, & Pilditch, 2018; Pilditch, Hahn, & Lagnado, 2018; Pilditch et al., 2019) set as its goal the development of assistive technology that could facilitate group decision-making in an intelligence context. To this end, BARD provides a graphical user interface enabling intelligence analysts to represent arguments as CBNs and allowing them to examine the impact of different pieces of evidence on arguments as well as to bring groups of analysts to a consensus via an automated Delphi method. An essential component of the system is the algorithm for generating natural language explanations of inference in a CBN, or more specifically, an explanation of evidence propagation in a CBN. This algorithm builds on earlier work by Zukerman and colleagues that have sought to use CBNs to generate arguments (Zukerman, McConachy, & Korb, 1998; Zukerman, McConachy, Korb, & Pickett, 1999). The algorithm uses an evidence-to-goal approach to generate explanations for a CBN. An explanation starts with the given pieces of evidence and traces paths that describe their influence on intervening nodes

until the goal is reached. In essence, the algorithm adopts a causal interpretation of the links between the connected nodes, finds a set of rules that describe causal relations in a CBN, and calculates all paths between evidence nodes and target nodes (claims/hypotheses) and builds corresponding trees in order to determine the impact of evidence on target nodes. Figure 3.4 provides an example. There we have four pieces of evidence: *Emerson report*, *Quinns report*, and *AitF Sawyer Report* all stating that 'The Spider' is in the facility and *Comms Analyst Winter Report* stating that 'The Spider' is not in the facility. The goal is to explain the impact of these four pieces of evidence on two target variables, namely *Is The Spider in the facility?* and *Are logs true? (Are Emerson & Quinn spies?)*. The algorithm first finds all relevant paths between evidence and target nodes, builds a corresponding tree and calculates the impact of evidence on the target, which is simply a difference between the probability of the target node *before* learning particular piece(s) of evidence and *after* learning particular piece(s) of evidence. This way the algorithm can find the so-called *HighImpSet*—the nodes that have the highest impact on the target node, the *NoImpSet*—the nodes that, in light of the other evidence nodes, have no impact on the target node, and the *OppImpSet*—the nodes that have the opposite impact to that of *HighImpSet*. Finally, the algorithm realizes the explanations in English language using sentences, clauses and phrases devised and combined by means of a semantic grammar (Burton, 1976). The output of the algorithm is presented in Figure 3.5.

As can be seen, the output in Figure 3.5 provides significantly more information to the user than just a single verdict on whether or not the variable *Is The Spider in the facility?* is part of the explanation of the evidence, as would be

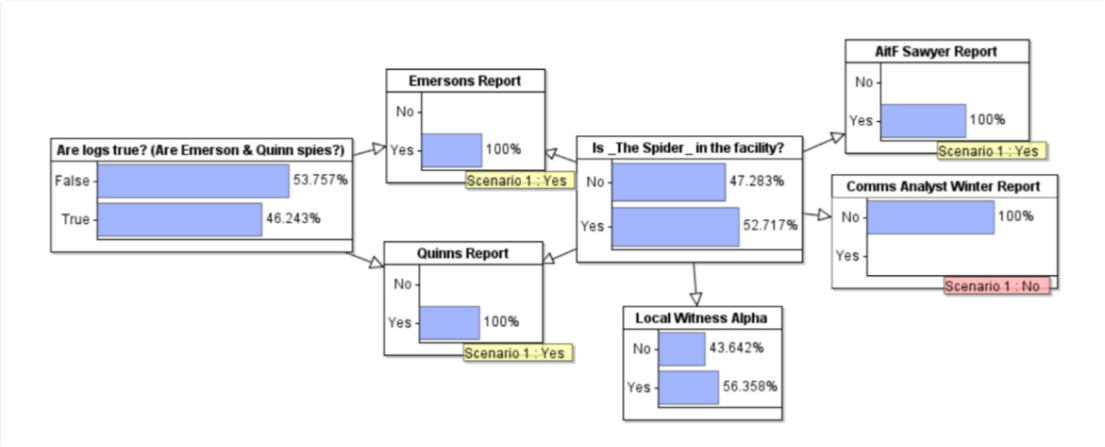


Figure 3.4: A CBN of a fictional scenario used in BARD testing phase. Four pieces of evidence are available: *Emerson Report*=Yes, *Quinns Report*=Yes, *AitF Sawyer Report*=Yes, and *Comms Analyst Winter Report*=No.

Is The Spider in the facility?		
HighImpSet	$\{\{ASR\}\}$	In the absence of evidence, the probability of *Is The Spider in the facility?* = Yes is 10% (very unlikely). Observing *Emerson's Report* = No and *Quinn's Report* = No reduces the probability of *Is The Spider in the facility?* = Yes. However, adding the evidence *AitF Sawyer's Report* = Yes increases the probability of *Is The Spider in the facility?* = Yes. The final probability of *Is The Spider in the facility?* = Yes is 5.3% (very unlikely).
MinHIS	$\{\{ASR\}\}$	
CombMinSet	$\emptyset$	
NoImpSet	$\emptyset$	
OppImpSet	$\{\{ER\},\{QR\}\}$	
Are the logs true? (Are Emerson and Quinn spies?)		
HighImpSet	$\{\{ER\}, \{QR\}\}$	In the absence of evidence, the probability of *Are the logs true?* = True is 10% (very unlikely). Observing either *Emerson's Report* = No or *Quinn's Report* = No reduces the probability of *Are the logs true* = True. However, adding the evidence *AitF Sawyer's Report* = Yes increases the probability of *Are the logs true?* = Yes. The final probability of *Are the logs true?* = Yes is 4.8% (almost no chance).
MinHIS	$\{\{ER\}, \{QR\}\}$	
CombMinSet	$\{ER, QR\}$	
NoImpSet	$\emptyset$	
OppImpSet	$\{\{ASR\}\}$	

Figure 3.5: A summary report generated by the BARD algorithm applied on the CBN from Figure 3.4. In addition to the natural language explanation, it provides sets with nodes that are *HighImpSet*, *NoImpSet* and *OppImpSet*. For the purposes of this chapter we can ignore *MinHIS* and *CombMinSet*.

the output of methods looking for a justification of evidence. In addition to the impact sets, it provides a natural language explanation on how different pieces of evidence influence the probability of the target variable (claim/hypothesis). Here again explanation can be explicated as the bridge or the link between evidence (data) and the target variables (claims/hypotheses).

Nevertheless, there remain continued challenges with this approach. First, the algorithm retains difficulties in coping adequately with soft evidence, namely evidence that we do not learn with probability 1. For instance, imagine that in the CBN in Figure 3.4 we additionally learn that the logs are most likely true, but we are not absolutely convinced. To reflect that we would set *p(Are logs true? (Are Emerson & Quinn spies)* = True) to equal 0.95 for instance. Thus the probability of *Are logs true? (Are Emerson & Quinn spies)* = True has changed from 0.46243 to 0.95, but it didn't go all the way to 1. The current version of the algorithm is not able to calculate the impact of such change. Second, the explanations generated by the algorithm are not aimed specifically at what a human user *might find hard to understand*. To make matters worse, it is arguably the interactions between variables and the often counterintuitive effects of these, that users will most struggle with (for psychological evidence to this effect see for example Dewitt et al., 2018; Liefgreen et al., 2018; Phillips et al., 2018; Pilditch et al., 2018, 2019; Tešić et al., 2020). In other words, the system generates an (accurate) explanation, but not necessarily a good explanation. For further guidance on what might count as a good explanation I consult research on this topic within the philosophy of science and epistemology, where the topic has raised decades of interest.

3.1.3 Good explanation

3.1.3.1 A brief overview of models of explanation

The historic point of departure for thinking about the nature of explanation in philosophy is the covering law model (Hempel & Oppenheim, 1948), also known as the "deductive-nomological model" of scientific explanation (where nomological means pertaining to the laws of nature). This model construes explanation as a deductive argument with true premises that has the phenomenon to be explained (the so-called explanandum) as its conclusion. Specifically, this conclusion is derived from general laws and particular facts. For example, an explanation of a position of a planet at a point in time consists of a derivation of that position from the Newtonian laws governing gravity (general law), and information about the mass of the sun, the mass of the planet, and position at a particular time and velocity of each (particular facts) (Woodward, 2017). A key feature of this model is that it views explanation and prediction as essentially two sides of the same coin. In the same way that Newtonian laws and information about the mass of the sun and the planet etc. can be used to predict the position of the planet at some future time the inference can also be used to explain the position of the planet after we observe it. In other words, we see here the same tight coupling between diagnostic reasoning and predictive reasoning that I mentioned earlier in the context of CBNs. However, while this coupling works in CBN's across the range of possible probabilities, it becomes forced in the covering law model when dealing with probabilistic explanations, in particular, when dealing with cases where the probability of observing the conclusion is low. Not only do probabilistic

contexts move the inference from deduction to an ampliative inference where the conclusion is no longer certain, the symmetry between explanation and prediction also becomes forced. We might for example readily explain someone being struck by the lightning by appealing to stormy weather conditions and the fact that they were out in the open. But we would nevertheless hasten to predict that someone will be struck by the lightning even if they are out in the open and there is a storm as it is a low probability event. This limits the utility of the covering law model within the social sciences where deduction is not commonplace and where low probability events are often found. Hempel himself was aware of these difficulties to the extent that he proposed two versions of the model the deductive-nomological model and an inductive-statistical one, and himself thought that the inductive statistical model applied only when the explanatory theory involves high probabilities. Even this restriction, however, does not deal appropriately with the asymmetries involved in explanation. These can be observed even in purely deductive context as is illustrated by the following example from Salmon (1992). Imagine there is a flagpole with a shadow of 20m and someone asks why that shadow is 20m long. In this context, it seems appropriate to explain the length of the shadow by appealing to the height of the flagpole, the position of the sun, and the laws of trigonometry. These together adequately explain the shadows length. But note that this inference can be reversed: when can also use the sun's position, the laws of trigonometry, and the length of the shadow to explain the height of the flagpole. This, however, seems wrong; an adequate explanation of the height of that flagpole presumably involves an appeal to the maker of the flagpole in some form or other. Examples such as these serve to illustrate not just

the limits of Hempel's account but of the limits of deductive approaches in the context of explanation more generally.

The asymmetric relations involved in explanation prompted alternative accounts of scientific explanation within the subsequent literature. Chief among these are causal accounts which assert that to explain something is to give a specification of its causes. The standard explication of cause in this context is that of factors without which something could not be the case (i.e. *conditio sine qua non*). This deals readily even with low probability events, and causes can be identified through a process of "screening off". If one finds that $p(M \mid N, L) = p(M \mid N)$, then N screens off L from M and that M is causally irrelevant to L. For example, a reading of a barometer (B) and whether there is a storm (S) are correlated. However, knowing the atmospheric pressure (A) will make these two independent: $p(B \mid A, S) = p(B \mid A)$, suggesting no causal relationship between B and S. However, the notion of cause in itself is notoriously fraught as is evidenced by J. L. Mackie's convoluted (Mackie, 1965) definition whereby a cause is defined as an "insufficient but necessary part of an unnecessary but sufficient condition". This rather tortured definition reflects the difficulties with the notion of causation when multiple causes are present thus giving rise to overdetermination (for example, decapitation and arsenic in the blood stream can both be the causes of death), the difficulties created by causal chains (for example, tipping over the bottle which hits the floor which releases the toxic liquid) and the impact of background conditions (for example, putting yeast in the dough causes it to rise, but only if it is actually put in the oven, the oven works, the electrical bills have been paid, and so on). It is a matter of ongoing research to what extent causal Bayes nets, that is CBN's

supplemented with the *do*-calculus (Pearl, 2009), provide a fully satisfactory account of causality and these difficulties (see also Halpern & Pearl, 2005a). At the same time, the difficulty of picking a single one out of multiple potential causes points to the second main alternative to Hempel's covering law model, namely so-called pragmatic accounts of explanation.

According to van Fraassen (1977) an explanation always has a pragmatic component: specifically what counts as an explanation in any given context depends on the possible contrasts the questioner has in mind. For example, consider the question "why did the dog bury the bone?". Different answers are required for different prosodic contours: "why did the *dog* (i.e., not some other animal) bury the bone?"; why did the dog *bury* the bone? (say, rather than eat it); why did the dog bury the *bone*? (say, rather than the ball). In short, pragmatic accounts bring into the picture the recipient of an explanation while rejecting a fundamental connection between explanation and inference assumed by Hempel's model.

3.1.3.2 Explanatory virtues

Philosophy has not only tried to characterise the nature of explanation, it has also sought to identify the so-called "explanatory virtues". Of the many things that might count as an explanation according to a particular theoretical account of explanation, not all may seem equally good or compelling. Among 'explanations', we might ask what distinguishes better ones from poorer ones. In search of explanatory virtues that characterise good explanation, a number of factors have been identified: explanatory power, unification, coherence, and simplicity are chief among these. Explanatory power often relates to the ability

of an explanation to decrease the degree to which we find the explanandum surprising; the less surprising the explanadum in light of an explanation the more powerful the explanation. For instance, a geologist may find a prehistoric earthquake as explanatory of deformation in layers of bedrock to the extent that these deformations would be less surprising given the occurrence of such an earthquake (Schupbach & Sprenger, 2011). Unification refers to explanations' ability to provide a unified account of a wide range of phenomena. For example, Maxwell's theory (explanation) managed to unify electricity and magnetism (phenomena). Coherence renders explanations that better fit our already established beliefs to be preferred to those that do not (Thagard, 1989). Explanations can also have internal coherence, namely how parts of an explanation fit together. Finally, an often motioned explanatory virtue is simplicity. According to Thagard (1978), simplicity is related to the size and nature of auxiliary assumptions needed by an explanation to explain evidence. For instance, the phlogiston theory of combustion needed a number of auxiliary assumptions to explain facts that are easily explained by Lavoisier's theory: it assumed existence of a fire-like element 'phlogiston' that's given away in combustion and that had 'negative weight' since bodies undergoing combustion increase in weight. Others operationalise simplicity as a number of causes invoked in an explanation: the more causes the less simple an explanation (Lombrozo, 2007).

While all of these factors seem intuitive, debate persists about their normative basis. In particular, there is an ongoing debate within the philosophy of science about whether these factors admit of adequate probabilistic reconstruction (Glymour, 2014). Wojtowicz and DeDeo (2020), however, aimed to provide a Bayesian account of explanatory virtues and operationalize these

virtues in a common mathematical framework. At the same time, there is now a sizeable program within psychology that seeks to examine the application of these virtues to every day lay explanation. This body of work probes the extent to which lay reasoners endorse these criteria when distinguishing better from worse explanations (Bechlivanidis, Lagnado, Zemla, & Sloman, 2017; Bonawitz & Lombrozo, 2012; Johnson, Jin, & Keil, 2014; Johnson, Johnston, Toig, & Keil, 2014; Lombrozo, 2007, 2016; Pennington & Hastie, 1992; Sloman, 1994; Williams & Lombrozo, 2010; Zemla, Sloman, Bechlivanidis, & Lagnado, 2017). To date, researchers found some degree of support for these factors, but also seeming deviations in practice.

3.1.3.3 Implications

What can be inferred for the project of explaining reasoning in CBNs and in AI systems in general from this body of research? There are at least two points.

First, it seems clear that CBNs provide a potential tool that is compatible with present thinking about the explanation at least in principle. They can capture the asymmetry in explanation as arcs are directed and can have a causal interpretation (Pearl, 2009), whilst at the same time being able to make predictions. This is in contrast to, for instance, a rule-based expert system with IF-THEN rules and a set of facts which would be susceptible to the symmetry 'error' in explanation illustrated by the flagpole example from Section 3.1.3.1. A CBN on the other hand would be able to account for the asymmetry given a causal interpretation and directional representation of arrows. However, it is neither clear how explanations in CBNs can capture the pragmatic component that van Fraassen raises nor how to operationalise explanatory virtues in the

context of CBNs. These are all potential avenues for further research.

Second, the debates about the nature of explanation and explanatory virtues have been conducted at very high levels of abstraction. They have also typically focused on philosophy of science and issues tightly related to it. This is true even for psychological research on explanation, to the extent that it has tried to model psychological investigations more or less directly on philosophical distinctions. However, for the purposes of developing suitable AI algorithms, it also seems important to work in the opposite direction, as it were from the bottom up. In other words, it seems important to simultaneously start with simple applications of CBN's to multiple variable problems, and consider what kinds of explanations a human (expert) would produce. This would shed light on the kinds of explanations that seem natural and appropriate to human users as well as provide guidelines on possible theories of explanation. A similar point has been made in AI literature that has emphasised the importance of human-generated explanations to serve as a baseline for comparison with machine-generated explanations (Doshi-Velez & Kim, 2017). To explore these ideas further, I conducted a case study on explanations of inference processes in CBNs.

3.2 A case study on human-generated explanation of inferences in CBNs

The main goal of this study was to explore the kinds of explanations of inference processes a human user may find appropriate and natural in the context of CBNs. This is interesting from both the psychological and the AI perspec-

tive as, on the one hand, it could give us further inputs into human explanatory intuitions and preferences and, on the other hand, it could inform the AI researcher that aims to build algorithms for an automated generation of explanations.

To explore this goal I have adopted a 'bottom up' approach. Namely, instead of assuming a particular definition of explanation of inferences in AI systems I ask a human user to provide us with explanations. I then analyze these explanations in the light of the above literature in philosophy, cognitive science, and psychology to come to a set of features that characterize explanations in AI systems which could then serve basis for XAI machine-generated explanations.

To the best of my knowledge, so far there is only one empirical study by Pacer, Williams, Chen, Lombrozo, and Griffiths (2013) that has compared human intuitions on explanations related to different causal structures to automated explanation based on four approaches to automated explanation in CBN discussed above, including MAP, MRE, and CET. They asked participants to provide (best) explanations of evidence (Experiment 1) or to rank explanations from best to worst (Experiment 2) and compared these results to outputs of the four approaches to automated explanation in CBNs. One of the findings was that human-generated explanations agree more with automated explanations that on some level include causal intervention. However, the study was looking at explaining a particular event (outcome) via another event (a node in a CBN) that (best) explains evidence. The question as to why participants choose a particular event as the best explanation or why they rank explanations (causes) the way they did is left unaddressed. In the terminology introduced in the first part of the chapter, Pacer et al. (2013) were exploring explanations

of outcomes in a CBN, rather than the explanations of reasoning processes in a CBN. In contrast, in what follows I focus on the explanations of inference processes in CBNs.

3.2.1 Overview

In the case study I explored how CBN human experts explain (diagnostic, predictive, and intercausal) inferences in CBN models. To do so I asked the experts to produce explanations of the probability change (or no change) of a target node (claim/hypothesis) in a CBN after learning particular evidence. The hope was that this would trigger the experts to provide explanations of *reasoning processes* in a CBN that lead to that change in the probability, rather than simply selecting a set of nodes that are responsible for this change. The experts answered questions related to the change in the probability in four different CBNs of different complexities.

3.2.2 Methods

3.2.2.1 Participants

I recruited three independent CBN experts who actively use CBNs in their research in computer science, cognitive psychology, and philosophy to participate in the case study. All experts have many years of experience in CBN modeling (the range is between 5 years and 40 years) and are well published in the literature on both theoretical foundations and applications of CBNs. The main reason for asking experts rather than non-experts (or both) to participate in the study is that CBNs are a technical tool that requires deep knowledge

and experience in order to understand the reasoning processes in them. Even though non-experts are able to use this tool with some success after a few hours of training (see Cruz et al., 2020), it is unlikely that the training provided to non-experts would provide them with the in-depth knowledge needed to explain sometimes complex reasoning processes in CBNs. Furthermore, all three experts have experience in communicating CBNs and their workings to the non-expert audiences, either through lectures, conferences, or collaborations in different academic and industry settings.

3.2.2.2 Materials

The three experts were presented with four well-known and publicly available CBN models ('Wet grass', 'Chest clinic', 'False barrier', and 'Car diagnosis') publicly available on an online CBN repository `https://www.norsys.com/netlibrary/index.htm`. The complexity of the four CBNs is varied: the number of nodes in the CBNs ranged from 4 to 18 and the number of arcs raged from 4 to 20. The four CBNs also differed in the target system they were modeling: three of the CBNs (Wet grass, Chest clinc, and Car diagnosis) had real-world target systems where variables in the models consisted of real-world events (such as *Rain*, *XRay test result*, *Battery voltage*); one CBN (False barrier) had a fully abstract target systemt where variables were denoted with letters from A to D.

There are two reasons for presenting participants with CBN models that they were already familiar with: (i) it is quite likely that the experts were previously asked to provide explanations of the reasoning processes in these four networks, which would imply that their answer were already rehearsed and

validated by their peers as well as non-expert audiences; and (ii) it would save time on the part of the participants to complete the questionnaire.

The experts were given four files implementing the fully parameterized four CBNs (in Netica and AgenaRisk formats). An example of the Wet grass CBN model implemented in AgenaRisk a can be found in Figure 3.6. The experts were told that they should be using the four models to answer the questions in the questionnaire that accompanied the files.

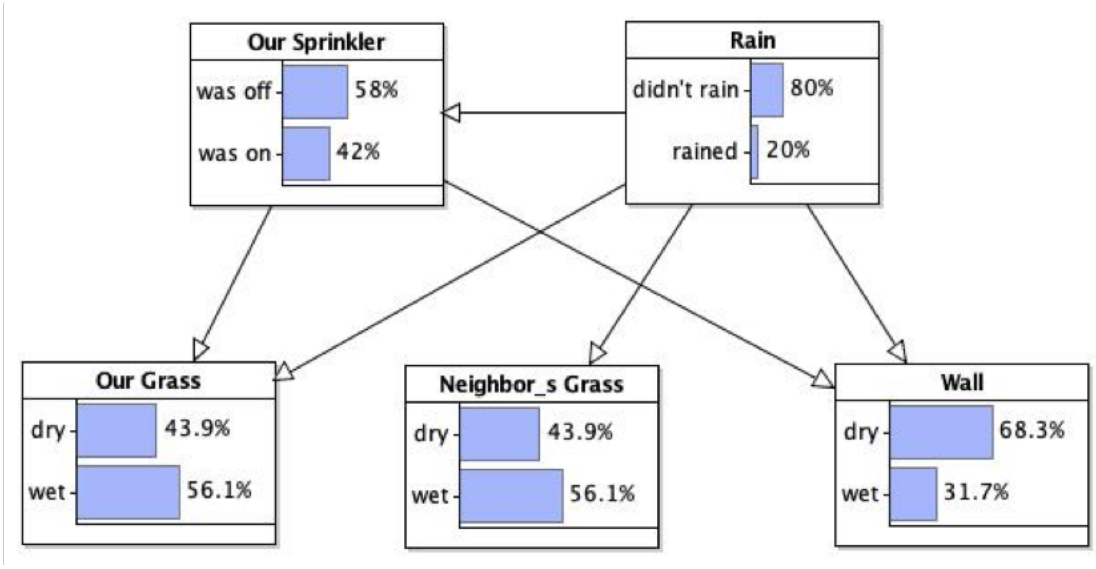


Figure 3.6: An AgenaRisk implementation of the 'Wet Grass' CBN model used in the case study.

The questionnaire consisted of 35 questions (28 required questions and 7 optional questions) regarding the inferences in the four CBNs. Out of these 35 questions, 8 required and 2 optional questions were about inferences in the Wet grass model, 6 required and 1 optional question were about inferences in the Chest clinic model, 6 required and 2 optional questions were about the False barrier model, and 8 required and 2 optional questions were about inferences

in the Car diagnosis model. The format of the questions was chosen such that it probes experts' intuitions on explanations of (diagnostic, predictive, and inter-causal) reasoning processes in CBNs. The questions prompted the experts to consider how learning evidence changed the probabilities of the query nodes in a CBN model. The general format was: given the evidence X, how does the probability of Y change compared to when that evidence was not available and why? The questions were also aiming to elicit explanations of different kinds of reasoning: 10 questions aimed at eliciting explanations of diagnostic reasoning, 7 questions aimed at eliciting explanations of predictive reasoning, 10 questions aimed at eliciting explanations of combinations of diagnostic and predictive reasoning, and 8 questions aimed at eliciting explanations of inter-causal reasoning. The full materials can be found in Appendix B.

3.2.2.3 Procedure

The three experts were asked to consider the four CBNs in the four files. They were told that the CBNs are already fully parameterized and that they should use them as such to answer the questions in the questionnaire. They were also told that the curly brackets, i.e. {}, indicate all evidence that they have available and that they are supposed to use only the evidence in {} to update the CBN model and answer the subsequent question. Further, they were told that some questions were optional, but that it was preferable that they answer as many questions as they could. Following this short introduction the experts were presented with the questions related to the four CBNs. For example, a question asked in relations to the Wet grass CBN model from Figure 3.6 was:

- Given evidence: {*Neighbors grass* = wet}

Question: How does the probability of '*Our Sprinkler* = was on' change compared to when there was no evidence and why?

This meant that the experts should input {*Neighbors grass* = wet} as evidence in Wet grass model and update the model to result in:

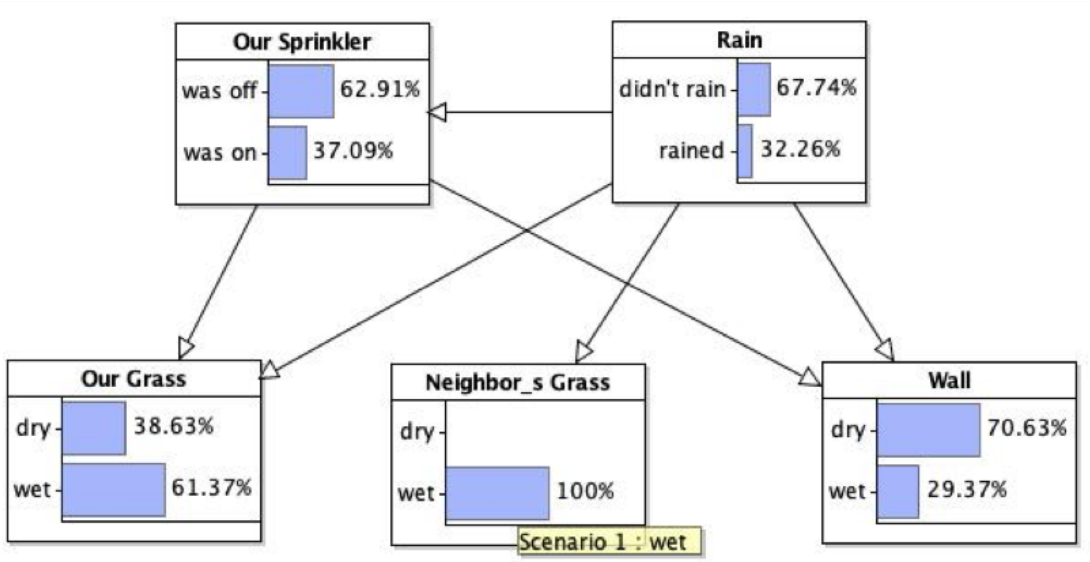


Figure 3.7: An AgenaRisk implementation of the 'Wet Grass' CBN model used in the case study when it is known that the neighbor's grass is wet.

Next, they should compare the probability of '*Our Sprinkler* = was on' when it *was not known* that {*Neighbors grass* = wet} (i.e. probability of '*Our Sprinkler* = was on' from Figure 3.6) to the probability of '*Our Sprinkler* = was on' when it *was known* that {*Neighbors grass* = wet} (i.e. probability of '*Our Sprinkler* = was on' from Figure 3.7) and write a short explanation as to why the probability changed.

3.2.3 Results and Discussion

Of the maximum 105 answers (35 per expert) to questions in the questionnaire 83 were received: 28 answers to question related to the Wet grass model, 20 related to the Chest clinic model, 15 related to the False barrier model, and 20 related to the Car diagnosis model. One expert provided answers to only the required questions regarding the Wet grass and Chest clinic models.

The three independent sets of answers were subjected to an analysis by a fourth person (also an expert in CBNs with decades worth of experience) in order to identify both commonalities and differences across the answers. This formed the basis of the subsequent evaluation of those answers with the following findings.

Firstly, I observed that the experts did provide explanations of reasoning processes in CBNs and did not simply point to a set of nodes in a CBNs that are 'responsible' for the change in probability. For example, in response to how and why the probability of '*Rain* = rained' changed in light of the new evidence Expert 3 answered:

> It [the probability] decreases from 20 to 10.77. Although Our Grass = wet increases probability of Rain = rained (to 32.26), Wall = wet reduces it because it strongly increases Our sprinkler = on, which in turn decreases Rain = rained more than Wall=wet increases it.

Furthermore, the explanation seemed to have played a role of a link (bridge) between the target node (claim/hypothesis) and evidence in a similar manner that, as we have seen above, the explanation 'A Swede is most likely not a Roman Catholic' connects the evidence that a person is a Swede to the claim

they are not Roman Catholic. This is an important result as it suggests that (i) human experts are capable of producing explanations of reasoning processes in CBNs and (ii) these explanation have the function of a link between the target node and evidence. It further suggests that these explanations could be used to inform XAI researchers who explore automated explanations of reasoning processes in AI systems.

Secondly, there was a high level of agreement across the experts' answers. Differences were typically more presentational than substantive. For example, the following three statements all seek to describe the same state of affairs:

- 'As A is true C is more likely to be true if B is true and less likely to be true if B is false. As I do not know B these alternatives essentially cancel themselves out and leave the probability of C unchanged.'

- 'It does not change. $P(\mathrm{C} \mid \mathrm{A})$ is equal to $P(\mathrm{C})$ if $P(\mathrm{C} \mid \mathrm{A}, \mathrm{B}) = P(\mathrm{C} \mid \sim\mathrm{A}, \sim\mathrm{B})$ and $P(\mathrm{C} \mid \mathrm{A}, \sim\mathrm{B}) = P(\mathrm{C} \mid \sim\mathrm{A}, \mathrm{B})$ (assuming $P(\mathrm{B}) = 0.5$), which here is the case.'

- 'According to model parameters: If A and B both true or both false, then C has probability .75. If A true but B false, or vice-versa, then C has probability .25. When I know A is true, and prior for B is 50%, there is a 50% that probability of C is 75% and a 50% that probability of C is 2%, therefore overall probability of C is 50%.'

Thirdly, causal explanations were prevalent. All three experts highly relied on causal explanations:

> The probability of rain decreases because, although the sprinkler

> and rain can both cause our grass to be wet, the wet wall is more likely to happen when the sprinkler is on rather than rain.

Notably, in appealing to causes, it is the most probable cause that seems to be highlighted as an explanation:

> There is a decrease [in probability] because the most likely cause of our grass being wet is the sprinkler and since the wall is dry the sprinkler is unlikely to be on.

As we seen above, pointing to most probable cause is the way the MAP approach selects explanations. However, in contrast to MAP, the expert also provided an explanation of how the most likely cause figures in the context of other relevant variables that can affect the change in the probability. So, simply pointing to a set of nodes is not sufficient to explain an inference process, although it can be a part of it. Furthermore, I found that causal language was more prevalent in explanations of inferences of the three CBNs that have real-world domains as target systems, i.e. Wet grass, Chest clinic, and Car diagnosis, than in explanations of inference of the False barrier. For example, some typical explanations in related to the False barrier CBN are:

> It [the probability] does not change. $P(\mathrm{C} \mid \mathrm{A})$ is equal to $P(\mathrm{C})$ if $P(\mathrm{C} \mid \mathrm{A}, \mathrm{B}) = P(\mathrm{C} \mid \sim\mathrm{A}, \sim\mathrm{B})$ and $P(\mathrm{C} \mid \mathrm{A}, \sim\mathrm{B}) = P(\mathrm{C} \mid \sim\mathrm{A}, \mathrm{B})$ (assuming $P(\mathrm{B}) = 0.5$)), which here is the case.

Fourthly, all experts appeal to *hypothetical reasoning* as a way of unpacking interactions of evidence variables:

> *Wall* = wet is a lot more likely if the sprinkler was on than if it rained (as a matter of fact, if it rained, the wall is more likely to be dry than wet). Since, *Our sprinkler* = was on went down, *Wall* = wet went down.

This is an important finding as it suggests that XAI should focus also on explaining different possible scenarios ('contrasts') in accounting for evidence and not just focusing on explaining what actually happened (see also Miller, 2019). Contrastive explanations, thus, seem natural (confirming van Fraassen's intuitions about explanation), and in a CBN context this emerges as discussion of behaviour under alternative, hypothetical evidence states.

Finally, these data seem to suggest that the structure of the CBNs is exploited in order to zero in on a subset of variables that will feature in an explanation. Specifically, explanations seemed to make use of the Markov blanket discussed above. In addition to Markov blanket, experts' descriptions mostly followed the direction of evidence propagation, i.e. followed the directed paths in a CBN:

> The probability of *Battery voltage* = dead increases because failure of the car to start could be explained by the car not cranking and the likely cause of this is a faulty starter system. A dead battery is one possible explanation for a faulty starter system.

This suggests that the explanatory virtue of 'simplicity' might, in a CBN context, be conceptualised in terms of a Markov blanket and path direction, giving some support for the EBI procedure which uses Markov nodes as featuring in an explanation of evidence.

In summary, I found that experts are capable of explaining inferences processes in CBNs and that explanation here often have the function of a link or a bridge between the target node (claim/hypothesis) and evidence. Further, multiple features of the philosophical, psychological, and cognitive science literature are reflected in these explanations: a focus on causal explanation for a probabilistic system; the directional nature of explanation (its asymmetry); indications of pragmatic sensitivity in that hypotheticals are used to express relevant 'contrasts'; and, an emerging notion of simplicity in the use of the Markov blanket. I also found that although existing XAI procedures for automated explanation do not fully account for explaining inference processes, some of them are a part of these explanations.

These results are still very much preliminary and further research is needed. However, they still provide an initial sense of the kinds of explanations a human (expert) may produce and, potentially, prefer in the context of explanations of reasoning in CBNs, and more generally in AI systems. More specifically, the case study makes a start at bridging between machine reasoning, and the philosophical and psychological literatures on what counts as 'good reasoning' by eliciting explanations by human experts. The work illustrates how concrete cases rapidly move discussion beyond abstract considerations of explanatory 'virtues' toward specific targets more suitable for emulation by machines. At the same time, this highlights the limitations of present algorithms for generating explanations from CBNs as we have seen is the case with the BARD project. Nonetheless, it provides concrete direction for future algorithm construction. And it indicates that bottom up approaches such as the one taken are informative and should be pursued further in future.

Needless to say, the present study represents a first attempt only, and future work expanding on it is required. Such work, should include larger samples of experts, although the degree of convergence observed makes fundamental disagreements seem less likely. Crucially, however, it should also further broaden the range of examples and network structures considered: it is here, that I most expect interesting features to have been missed.

3.3 Conclusions

As AI systems come to permeate human society, there is an increasing need for such systems to explain their actions, conclusions, or decisions. This is presently fuelling a surge in interest in machine-generated explanation. However, there are not only technical challenges to be met here; there is also considerable uncertainty about what suitable target explanations should look like.

In this chapter I have presented two ways to understand different notions of what counts as an explanation. One of these notions involves explanation as identification of the variables that mattered in generating a certain outcome. In the context of computational models of explanation in CBN's this corresponds to the usual focus on explaining observed evidence via unobserved nodes within the network. In other words, the explanation identifies a justification/hypothesis. This is the notion of explanation that has figured prominently in work on computer-generated explanations as well as in psychological and philosophical literature on explanation. The second notion of explanation I considered includes explanation of the inference that links evidence and hypothesis. In the context of CBN's this means explaining the reasoning processes that lead to a change (or no change) in the probabilities of the query

nodes. In other words, the explanation of a target node (claim/hypothesis) involves information about the incremental reasoning process that identifies that hypothesis. It is this second notion of explanation that is crucial in explaining decision-making processes in AI systems that would arguably increase the transparency of and trust in the outputs of these systems. Explanation so understood constitutes a fundamental problem of human computer interaction and only empirical research that seeks to understand the human user can lead to fully satisfactory answers. Nonetheless, the empirical exploration of this second notion of explanation is lacking.

The case study from this chapter aimed at filling in the gap in the empirical exploration of the second notion of explanation. The findings of the case study, although limited, suggest that people (experts) are able to provide explanations of the reasoning processes in CBNs. These explanations seem to have many features found in the philosophy and psychology literature on explanation. However, some of these features such as the explanatory virtue simplicity seemed to have been operetionalized to fit the context of CBNs. This suggests that while the literature from philosophy and psychology is a helpful starting point, a more context specific work would be required to address the question of what would be to most appropriate explanation of the reasoning processes in a particular AI domain.

The two notions of explanation discussed in this chapter do not, however, exhaust all aspects of explanation. Explanations often include an explainer (a human or an AI system providing an explanation) and explainee (a person receiving an explanation); in other words, explanations are also communicative acts. The next chapter explores some of the implications of understanding ex-

planations as communicative acts.

4

Social explanation: The effects of explanation on reliability and confidence

The explanations studied in the previous two chapters were of an intrapersonal character. Namely, in Chapter 2 participants were asked to estimate an impact

of an explanation (a variable in a CBN) on occurrence of another event (another variable in a CBN) and in Chapter 3 they were asked to to provide an explanation of reasoning processes in a CBN that they themselves would deem explanatory. The potential social dimension of explanations was set aside and the participants were asked to judge the impact of explanations or provide explanation that they would judge as good from their own, intrapersonal perspective.

In Chapter 1, however, we have seen that arguments have a clear social aspect. They often involve providing a (rational) support for or against a position in order to persuade other people. If arguments are in many ways comparable to explanations, and explanations themselves might sometimes function as arguments, then it is plausible and worth exploring the interpersonal context of explanations. Indeed, even prima facie it seems that explanations do have a social aspect as well. They often include at least two parties: an explainee, a person who is receiving an explanation and an explainer, a person (or sometimes a machine) who is providing an explanation. For example, virtually all education settings involve a teacher who sometimes provides explanations to the students with the goal to increase the students' understanding of the topic. Experts often provide explanation on the their topics of expertise to both the other experts and non-experts alike. An AI system may provide an explanation of its decision-making processes to human users.

Some of the implications of considering arguments in a social context are the newly emerging factors that may influence the strength of the arguments. One of these factors, as pointed in Chapter 1, is the source's reliability. The research on source reliability in the context of argumentation highlighted the

interplay between the content of an argument that the speaker is putting forward and the speaker's reliability. This interplay can only be explored if we consider arguments as having a social dimension. Drawing parallels with arguments, one would then expect that once we start exploring explanations in a social context we would also encounter factors, such as the reliability of an explainer, that would significantly influence the impacts of explanations. In this chapter, I thus experimentally investigate some of these factors. In particular, I explore the potential effects of the reliability of an explainer on the beliefs of the recipient of an explanation, i.e. an explainee.

An explainer, however, can provide different types of explanation. In the previous chapter we have seen that there are at least two different notions of explanation. Thus, one could study the effects of the explainer's reliability with respect to either notion. In this chapter I focus on one of these notions: namely, explanations of reasoning processes. More specifically, I explore the effects of the reliability of an explainer and the explanations understood as the links or bridges between hypotheses/claims and evidence on explainee's beliefs. In the context of the 3-dimensional explanation cube, these explanation are still products in a sense that they are results of an explanation processes, they are made in an interpersonal context, and they are explanations of reasoning processes (see Figure 4.1).

The chapter proceeds as follows. First, I briefly summarize the notion of explanation explored in this chapter, discuss the potential effects of explanations, and the implications of their social character such as the reliability of an explainer. I then present four experiments that test the effects of these explanations and the reliability of an explainer on the beliefs of the explainee.

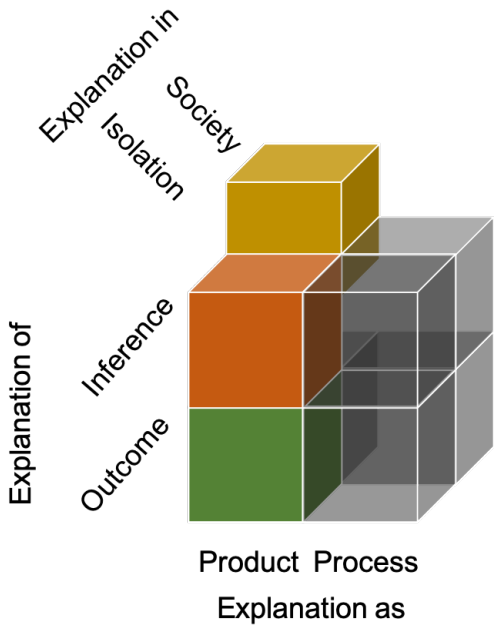


Figure 4.1: The three dimensions of explanation.

4.1 Introduction

4.1.1 Explanations: connecting claims with evidence

In the previous chapter we have seen that one can distinguish between at least two notions of explanation found in the literature in philosophy, psychology, and computer science: explanations of evidence (outcome) and explanations of reasoning processes. Although empirically under-explored, the case study from the previous chapter suggested that people are capable of providing explanation of the reasoning processes at least in the context of CBNs. Furthermore, the explanations that the experts who participated in the case study generated had a function of a link or a bridge between target nodes (hypotheses)

and evidence, explaining how the evidence lead to a change in the probability distribution of a target node.

In the previous chapter I have also argued that the view of explanations as links between hypotheses (claims) and evidence (data) is not novel and that it can be found in the argumentation literature as well as in the computer science literature, in particular the literature pertaining to recommender systems and expert systems. Some of the examples illustrating this included the explanation 'A Swede is most likely not a Roman Catholic' that elucidates the relationship between for example the evidence that a person is a Swede and the claim they are not Roman Catholic. In this chapter, I will focus on explanations of reasoning processes, in particular, those that have a function of connecting or elucidating the connection between claims and evidence.

4.1.2 Effects of explanations

The effects of providing an explanation understood as a link between the hypothesis and data or explanations of reasoning processes has generally been under-explored in psychology and philosophy. In contrast, the effects of explanations of evidence (outcome) have been extensively explored and it is plausible to think that some of these effects translate for the explanations understood as links. Here I present a brief overview of these effects.

In Chapter 1 I have mentioned that the main goal of explanation is increasing understanding. More precisely, providing an explanation of a phenomenon may increase the recipient's sense of understanding of what has been explained (Hempel, 1965; Hahn, 2011; Lombrozo, 2012). However, this increase in the sense of understanding may not result in an increase in actual understanding

as has been pointed out by Trout (2002, 2008). Further, it has been found that generating explanations (even incorrect ones), rather than just receiving explanations, has a beneficial impact on learning (Lombrozo, 2012). Explanations can also increase the perceptions of normality: finding plausible explanations of, for instance, patients' behaviour lead to the perception of the patients as being more 'normal' than when such an explanation was lacking (Ahn, Novick, & Kim, 2003). It has also been found that providing an explanation of a hypothetical outcome or of a past event that we are not sure if it had happened increases the likelihood of the hypothetical outcome to occur in the future and of the event that might have occurred in the past (Koehler, 1991, 1994; Ross, Lepper, Strack, & Steinmetz, 1977; Sherman, Zehner, Johnson, & Hirt, 1983).

These are only some of the effects of explanations. However, the effect of explanations that I am going to be focusing in this paper is the effect they have on the recipients' confidence in the claims/hypotheses. I review the literature regarding this effect next.

4.1.2.1 Effects of explanation on confidence

Inference to the best explanation (IBE) nicely illustrates the way the explanations impact confidence. The fact that some hypothesis or a claim is the best explanation (i.e. it has the highest explanatory goodness compared to the other rival hypothesis) increases the subjective probability (or confidence) assigned to that hypothesis. Douven and Schupbach (2015) suggest this is empirically also the case and that people judge a hypothesis more likely (i.e. they are more confident in the hypothesis being true) if the explanatory goodness of that hypothesis also increases. Some of the factors that influence the ex-

planatory goodness of a hypothesis and thus increase its subjective probability are simplicity (Lombrozo, 2007; Lagnado, 1994; Read & Marcus-Newhall, 1993; Thagard, 1978), breadth (Lombrozo, 2016; Read & Marcus-Newhall, 1993; Thagard, 1989), consistency with prior knowledge (Thagard, 1989) and coherence (Pennington & Hastie, 1993; Thagard, 1989) as discussed in the previous chapter.

Outside the context of IBE, and closer to the notion of explanation as links between claims and data, it has been found that asking people to provide an explanation as to whether a particular property is true or false changes their perceived likelihood of that property (Lombrozo, 2006). For instance, when asked to explain the relationship between two variables A and B (e.g. why risky people (A) are better firefighter (B)) participants' subjective estimates of the relationship significantly increased compared to both the control who was not prompted to explain the relationship (C. A. Anderson, Lepper, & Ross, 1980) and participants' previous estimates when they were not asked to explain the relationship (C. A. Anderson & Sechler, 1986). Here we again have a distinction between a hypothesis or a claim (e.g. high risk takers make better firefighters) and explanations that are provided in support for that claim (e.g. risky people act spontaneously and because speed is essential in fighting fires these kinds of firefighters are more successful) that increases people's confidence in the claim compared to the situation where people were not asked to provide an explanation (see Koehler, 1991).

Thagard (1989) similarly argues that if we are aiming to explain evidence (data) by arguing that the accused murdered the victim (claim or hypothesis), the hypothesis will be more plausible if we find reasons why the accused was

motivated to kill the victim (explanation). Here we again see how explanation plays a part in connecting data (evidence) with the hypothesis and how finding such a connection may result in the increased confidence in the hypothesis. Pennington and Hastie (1993) have empirically explored this idea. They find that the story summary (explanation), which is the interpretation of the evidence (data) that have a narrative story form, has an impact on the confidence in a juror's decision: the better the story (explanation) the greater the impact on the confidence. Finally, Brem and Rips (2000) also argue that that the perceived probability of the claim may be increased as a result of there being an explanation for that claim.

In summary, both the theoretical and experimental works suggests that ether providing or receiving an explanation will result in the increased confidence in the claim across different contexts and notions of explanation.

4.1.3 Explanations as communicative acts

Like arguments, explanations also have an important social dimension. They are often between individuals who try to communicate understanding (Keil, 2006), and they usually take the form of a conversation where "*[s]omeone* explains *something* to *someone*" (Hilton, 1990, p. 65, original emphasis). Explanations are then in their essence communicative acts (as highlighted by van Fraassen's pragmatic account discussed in the previous chapter) and, as such, involve interpersonal exchange and include two parties: an explainer and an explainee. In line with the literature mentioned in Section 4.1.2 one would then expect that the explainee's confidence in a claim would be affected by explainer's explanation. More specifically, one would expect that the act of the

explainer providing an explanation would increase explainee's confidence in the claim being explained.

The communicative dimension of explanations, however, introduces additional factors that could affect confidence. We often rely on others (e.g. experts) to provide us with explanations regarding some phenomena. For example, experts are called upon to explain to the general public why a particular virus is dangerous to the population. The fact that experts are providing us with an explanation may affect our confidence in the claim that the virus is dangerous. Now, explanations being communicative acts implies that they introduce information about the speaker (the explainer) that could affect the explainee's confidence. For example, one of the aspects of the explainer's that could affect explainee's confidence is the explainer's reliability.

The effects of the reliability of the source of information have been extensively explored, both theoretically and empirically, in the context of argumentation. Here, formal models of source reliability that aim at distilling the impact of reliability on confidence that goes beyond the argument content have been proposed by Bovens and Hartmann (2003) and Olsson and Vallinder (2013) (for a detailed review see Merdes, Von Sydow, & Hahn, 2020). Some of these models have been empirically tested. For instance, Hahn, Harris, and Corner (2009) varied both argument strength and the reliability of the sources and found that both argument strength and the reliability of the source affected the participants' confidence in the arguments, with an interaction between the two, which was in line with some of the formal models. Similarly, Harris et al. (2016) find that greater expertise and reliability increase the impact on one's confidence in claims (see also P. Collins, Hahn, von Gerber, & Olsson, 2018; P. Collins &

Hahn, 2019; Hahn et al., 2013; Hahn, Harris, & Corner, 2016; Walton, 2007). What is more, Jarvstad and Hahn (2011) find that perceived reliability can be affected by evidence (data) or the report from the source, with a more likely statement being judged to come from a more reliable source.

In this chapter I aim to explore the impact of the explainer's reliability on confidence in the claim both when there is an explanation for the claim and where such an explanation is missing. In line with the argumentation literature on argument content and reliability, one would expect to find differing impacts of reliability when an explanation for the claim is provided compared to when no explanation is provided.

To the best of my knowledge, the explanation literature not has experimentally manipulated the impact of reliability on confidence in claims to explore its impact. However, some limited exploratory analyses have been done. For instance, Zemla et al. (2017) explore criteria that predict explanation quality and find that expertise is one of the criteria that significantly predicts explanation quality, with the higher (perceived) expertise leading to a better quality of explanations. This potentially suggests that expertise positively impacts confidence: the more the explainer is perceived as an expert the higher the confidence in the claim. I aim to experimentally explore not only the impact of the explainer's reliability on the confidence in the claims, but also how providing an explanation and reliability combine to impact the confidence the claims.

4.1.4 Everyday explanations

Before I go on to describe the experimental exploration, a brief note on explanations used for this exploration. The philosophical literature on explanations

has mainly focused on scientific explanations. In Chapter 3 we have seen that the general motivation was to find what makes a good (or bad) explanation in science. The psychological investigations on explanations have mainly derived from the philosophical literature and studied aspects of explanations that the philosophical literature has considered to be important in judging scientific explanations (e.g. Lombrozo, 2007). Further, these empirical studies have too often employed short and simple explanations with a minimal causal structure with, sometimes, a single cause and effect (for an overview see Lombrozo, 2012).

More recently, however, psychologists have looked into everyday explanations to explore the sets of criteria that have been used to judge the explanatory goodness of these kinds of explanations (e.g. Bechlivanidis et al., 2017; Zemla et al., 2017). The aim of these studies was to explore whether the set of criteria for the goodness of the scientific explanations also played a role in the case of everyday explanations. Furthermore, exploring the explanatory criteria in the case of the real-world explanations provided a more ecologically valid understanding these criteria. Everyday explanations are more nuanced than the experimental stimuli often used in the psychological studies. As in this study I aim to explore the impact of explanations on the claim that go beyond the immediate impact of the data, everyday explanations are a more suitable experimental material for such exploration. Further, everyday explanations are very easily immersed in the conversational form where the communicative aspects of explanations and the impact of reliability on the confidence are more naturally explored. In this chapter, I thus use everyday explanations as materials for the empirical investigation of the impact of explanations and reliability.

4.2 Overview of experiments

The aim of this chapter was to explore the relationship between (everyday) explanations understood as links between claims and data, the reliability of an explainer, and the explainee's confidence in a claim. Experiment 6 tested the impact of explanations on the confidence in claim without any considerations of the social aspects of explanations, such the reliability of the explainer. The goal of this experiment was to replicate the findings from the previous literature regarding the impacts of explanations on confidence. Experiments 7a and 7b included the social aspects of explanation and tested the impact of explanations not just on confidence of the claim, but also on the reliability of the explainer. The aim here was to explore whether providing an explanation affects the reliability of the explainer. Experiment 8 explored the impact of both the explanation and reliability on the explainee's confidence in a claim, aiming to investigate the potential causal impact of the explainer's reliability on the confidence in a claim.

4.3 Experiment 6

The aim of Experiment 6 was to replicate the findings of previous studies on the effects of explanation on people's confidence using real-world explanations as stimuli. Following these studies, I expected that adding an explanation would increase people's confidence in the hypothesis.

4.3.1 Methods

4.3.1.1 Participants and Design

A total of 130 participants ($N_{\text{FEMALE}} = 87$, $M_{\text{AGE}} = 33.8$ years) were recruited from Prolific Academic (www.prolific.ac). All participants were native English speakers currently residing in the UK, the US, or Canada whose approval ratings were 95% or higher. They all gave informed consent and were paid £5 an hour rate for partaking in the present study, which took on average 10.5 min to complete.

Participants were randomly assigned to either the control group where no explanation of the claim was provided ($N = 66$) or the treatment group where an explanation was provided ($N = 64$).

4.3.1.2 Materials

Previously (see Section 4.1.4) I argued for the suitability of everyday explanation in exploring the impact of explanations on the confidence in the claims. Thus, in all experiments in this paper I used the following scenarios adapted from Zemla et al. (2017), who used Reddit's *Explain Like I'm Five* (Eli5; www.reddit.com/r/explainlikeimfive), Wikipedia, and HowThingsWork.com to source these stimuli. These platforms are widely accessible to the general population and the issues addressed on these platforms are often aimed at the general population, covering a wide range of phenomena that one can encounter in a daily life. The scenarios were picked from three different domains: public health, social policy, and history. The scenarios were chosen with the idea that general public would be interested in them. All scenarios had the

same format. The first paragraph started with an introduction of up to two sentences describing data/evidence (or sometimes referred to as explanandum). This was followed by a question seeking an explanation for the explanandum. The second paragraph described a claim that is supposed to account for the explanandum (the no explanation condition) or it described a claim and an explanation that accounted for the explanandum (the explanation condition). Lastly, participants were asked a question that elicited their confidence estimates in the claim. This format was very similar to the one adopted by Brem and Rips (2000, Experiment 2).

For example, *the Black Death* scenario looked as follows. Note that the text of the scenarios and the questions were the same for both the no explanation and the explanation conditions, except for the part in the parenthesis that appeared only in the explanation condition. The text in the square brackets did not appear in either condition and is added here to point to the functions of the different parts of the scenario.

> Millions of people died from the Black Death in the 14th century. [*data*] **How did the Black Death come to an end?** [*a prompt for an explanation*]
>
> One popular belief is that the Black Death subsided mostly through the use of quarantines. [*claim*] (According to this belief, people mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources would traditionally escape to the country, far away from the Black Death-infested cities. [*explanation*])

> **Q**. How confident are you that the Black Death came to an end through the use of quarantines? [*a question eliciting participants' confidence in the claim*]

The other four scenarios were concerned with the increase in China's population despite the one-child policy, the way medical practitioner contract Ebola, Switzerland's armed neutrality during World War II, and the way vaccines build immunity.

Zemla et al. (2017) experimentally studied the quality of explanations by asking participants to rank the explanations on the 7-point Likert scale from '1–Strongly disagree' to '7–Strongly agree' on how 'good' was the explanation. For all five explanations used in the five scenarios they found that participants rated them well above average in quality: the explanation from the Switzerland scenario had an average rating of 5.4, from the Ebola scenario 6.4, form the China scenario 5.1, from the Vaccination 5.7, and from the Black Death scenario 5.9.

For the full materials used in Experiment 6 see Appendix C.1.

4.3.1.3 Procedure

After giving an informed consent and basic demographic information, participants were shown the following instructions:

WELCOME!

You will now be presented with 5 explanations of 5 events and phenomena found in the real world and required to answer some questions related to the explanations.

Please make sure you read all the information carefully before answering the questions.

After these instructions, participants were presented with the five scenarios and questions related to these scenarios. The order in which the scenarios were presented was randomized for each participant. Each scenario was presented on two pages. On the first page was the main text of the scenario. On the second page, the text of the scenario was repeated as a reminder and participants were asked two questions: one about their confidence in the claim and another (which was the same for all scenarios) to explain their reasoning regarding how they arrived at their confidence estimate. The second question was asked to gain additional insight into participants' reasoning.

For example, *the Black Death* scenario had the following text on the first page (the additional text that appeared only in the explanation condition is in parenthesis):

> Millions of people died from the Black Death in the 14th century.
> **How did the Black Death come to an end?**
>
> One popular belief is that the Black Death subsided mostly through the use of quarantines. (According to this belief, people mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources would traditionally escape to the country, far away from the Black Death-infested cities.)

On the second page, the scenario was repeated and the questions related to

the scenario were asked:

> **Reminder:**
>
> Millions of people died from the Black Death in the 14th century. **How did the Black Death come to an end?**
>
> One popular belief is that the Black Death subsided mostly through the use of quarantines. (According to this belief, people mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources would traditionally escape to the country, far away from the Black Death-infested cities.)
>
> ***Please answer the following question.***
>
> **Q**. How confident are you that the Black Death came to an end through the use of quarantines?
>
> [A slider eliciting confidence (%) on a scale from 0% to 100%.]
>
> **R**. Please explain your reasoning for your answer to the question in the box below.
>
> [A text box.]

A percentage scale from 0% to 100% was used to elicit participants' confidence estimates in the claims in question. A free format type text box was used to ask participants to explain their reasoning for the estimates they provided. Lastly, after the participants answered questions to all five scenarios they received a debriefing information.

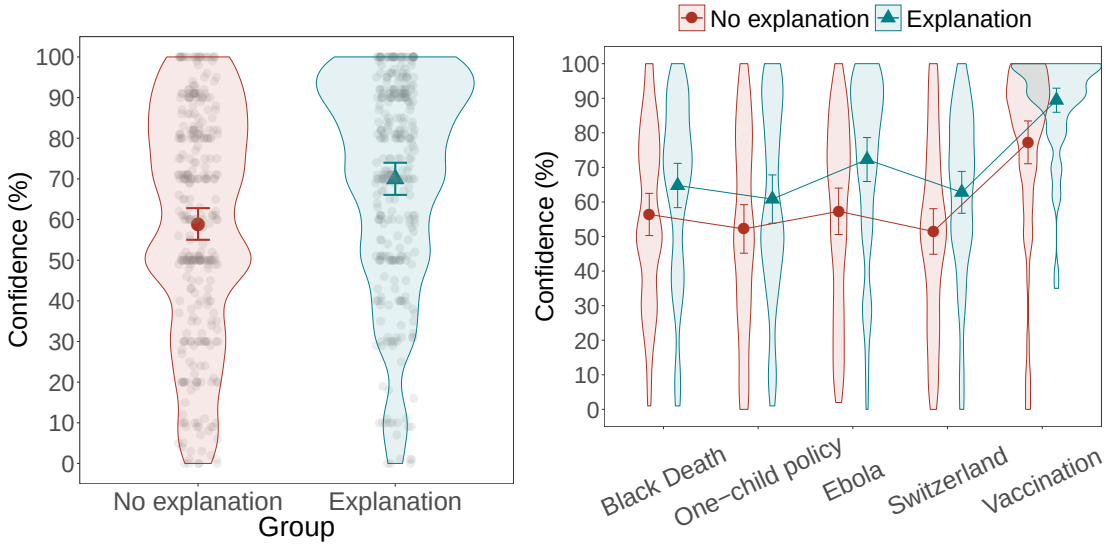
4.3.2 Results and Discussion

To analyze the data[1] I built a linear mixed effect model (LMM) using the lme4 package in R (Bates et al., 2014). The only fixed effect was group (with two levels: no explanation and explanation). The only random effect was the intercept for participants. There was no random slope from the participant as the design was fully between. No random intercept for scenarios was used as the number of scenarios was low (i.e. 5) and including the scenarios as a random intercept could have lead to a reduced power of the experiment (see Judd, Westfall, & Kenny, 2017; Singmann & Kellen, 2019). Further, a random slope for scenarios was not included as led to a singular fit model, implying that the variance of this random effect was (close to) zero.

The LMM indicated that confidence estimates in the explanation group (Estimated Marginal Mean $= EMM = 70.02$) was significantly higher than in the no explanation group ($EMM = 58.91$); $t(128) = 3.96$, $p < .001$ (see Figure 4.2a). Further, the inclusion of the predictor for the group in the model led to a significant improvement in model fit ($\chi^2(1) = 15$, $p < .001$), compared to just having an intercept as a predictor. This result is in the line with the previous literature and provides further support that people's confidence in claims is higher when explanation is provided.

The effect of explanation was not only observed overall, but also within each of the five scenarios. Figure 4.2b shows that participants' confidence estimates in the explanation conditions were higher in all scenarios, suggesting that the effect was not driven by specific scenarios. Further, the mean confi-

[1]Given the COVID-19 pandemic, it is worth mentioning that the data for all four experiments were collected in the period between June and September 2019.



(a) (b)

Figure 4.2: (a) The estimated marginal means (EMMs) from the LMM built for Experiment 6 with 95% confidence intervals. Gray points are raw data values (jittered along the x-axis for visibility) with violin plots showing the frequency of the raw data. (b) The observed data means (with 95% confidence intervals) and violin plots for each scenario broken down for each explanation conditions.

dence estimates were similar across the cover stories (in the respective explanation/no explanation conditions), expert in *Vaccination* scenario (particularly in the explanation condition of that scenario). From participants' textual answers where they provided reasons for choosing a specific confidence estimate in this scenario I noticed that a number of participants have said that the claim and explanation agreed with what they already knew about vaccination, which led them to provide higher estimates in both the no explanation and explanation conditions of this scenario. This finding hints at the importance of the background knowledge in judging people's confidence in claims supported by explanations. In the next experiments I find further support for the effects of

background knowledge.

4.4 Experiment 7a

In Experiment 6 participants were asked to provide their confidence estimates in claims in situations where the source of these claims is left out. However, as explanations are communicative acts they often include a speaker (an explainer) and as such could plausibly provide information about the speaker's reliability. The explanations then not only have an effect on confidence in the claims but they could also have an effect on the perceived reliability of the speaker who is providing an explanation.

The aim of this experiment was to explore the impact of an explanation on both the confidence in a claim as well as the perceived reliability of the source that provided the claim and explanation.

4.4.1 Methods

4.4.1.1 Participants and Design

A total of 52 participants ($N_{\text{FEMALE}} = 31$, $M_{\text{AGE}} = 32.6$) were recruited from Prolific Academic (www.prolific.ac). The selection criteria and the remuneration rate per hour were the same as in Experiment 6. Participants took on average 16.4 min to complete the experiment.

Participants were randomly assigned to either the control group where no explanation for the claim was provided ($N = 24$) or the treatment group where an explanation was provided ($N = 28$). All participants were asked to provide estimates regarding two dependent variables: confidence and reliability.

4.4.1.2 Materials

To explore the communicative aspect of explanations and their impact on the explainer's reliability I follow Hahn et al. (2009) who have studied the impact of reliability and the content of an argument on confidence in what is argued for. This is also in line with Walton (2004b) who argues that a dialogue from is an appropriate one for explanations.

The same five scenarios from Experiment 6 have been employed in this experiment and further adapted to fit the form of a dialogue between two people, an explainer and an explainee, where the explainer provided the claims and explanations. Such a format enables us to elicit not only participants' confidence estimates in claims but also their reliability estimates in the explainer as a source of the claims and explanations.

The adaptation of scenarios into dialogues was done in a similar manner as in Hahn et al. (2009). For example, *the Black Death* scenario was adapted in a way that it includes two people, an explainer and an explainee, where the explainee (Jimmy) is asking questions and the explainer (Dave) is trying to provide answers (the part in parenthesis appeared only in the explanation condition):

> Dave and Jimmy are part of a research group investigating devastating pandemics in human history. During a planning meeting they touched upon the Black Death.
>
> **Dave**: Millions of people died from the Black Death in the 14th century. I think our research project should in part focus on how the Black Death ended. It may give us some insight into how to deal

with future pandemics.

Jimmy: Yes, I agree. Do you already have an idea regarding how the Black Death came to an end?

Dave: I think the Black Death subsided mostly through the use of quarantines.

(**Jimmy**: How so?

Dave: People mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources would traditionally escape to the country, far away from the Black Death-infested cities.)

The other scenarios were adapted in a similar way. For full materials see Appendix C.2.

4.4.1.3 Procedure

The procedure for this experiment was similar to the procedure for Experiment 6 in that the welcome page was shown after the participants gave informed consent and demographic information, and each scenario was presented in a random order on two pages. The difference lies in that participants now answered two questions in each scenario: one about the confidence in the claim and one about the reliability of the explainer. For example, after being shown and reminded of *the Black Death* scenario the participants were asked:

Q1. How confident are you that the Black Death came to an end through the use of quarantines?

[A slider eliciting confidence (%) on a scale from 0% to 100%.]

Q2. How reliable do you think **Dave** is as a source of information regarding the end of the Black Death?

[A slider eliciting reliability (%) on a scale from 0% to 100%.]

For both the confidence questions and the reliability questions participants were asked to move the slider which was on the scale from 0% to 100%. Both the confidence and the reliability questions were followed by free format type text boxes where participants could explain their reasoning for selecting certain confidence/reliability estimates. Finally, participants received debriefing information.

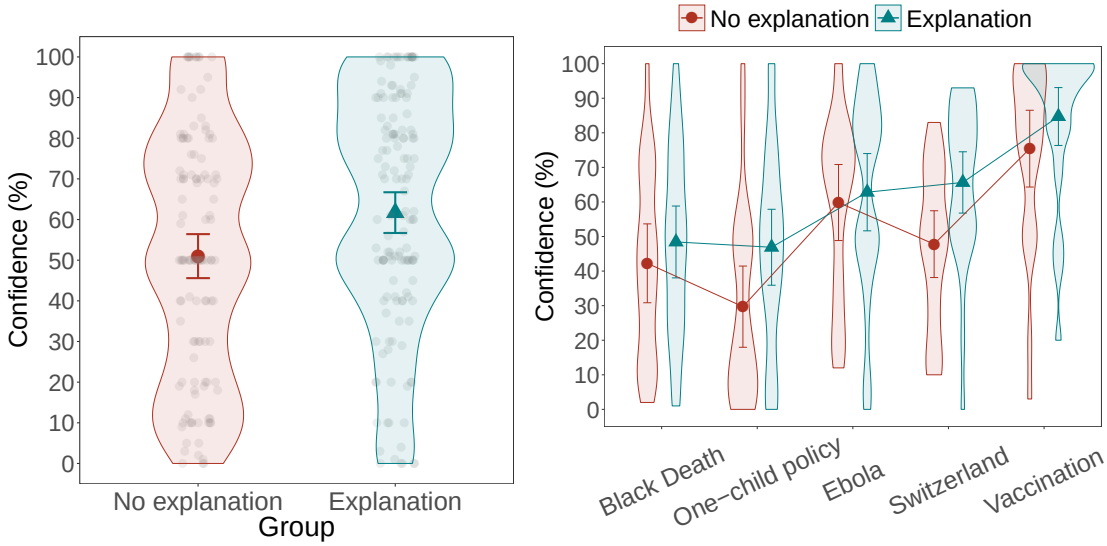
4.4.2 Results and Discussion

Separate analyses were conducted for each dependent variable.

4.4.2.1 Confidence

The LMM with the same random effects structure as in Experiment 6 indicated that confidence estimates in the explanation group ($EMM = 61.7$) were significantly higher than in the no explanation group ($EMM = 51$); $t(50) = 2.91$, $p = .005$ (see Figure 4.3a). Further, the inclusion of the predictor for the group in the model led to a significant improvement in model fit ($\chi^2(1) = 8.1$, $p = .004$), compared to just having an intercept as a predictor.

This is in line with the finding from Experiment 6 and the previous literature. The similar trend was found in all five scenarios (see Figure 4.3b).



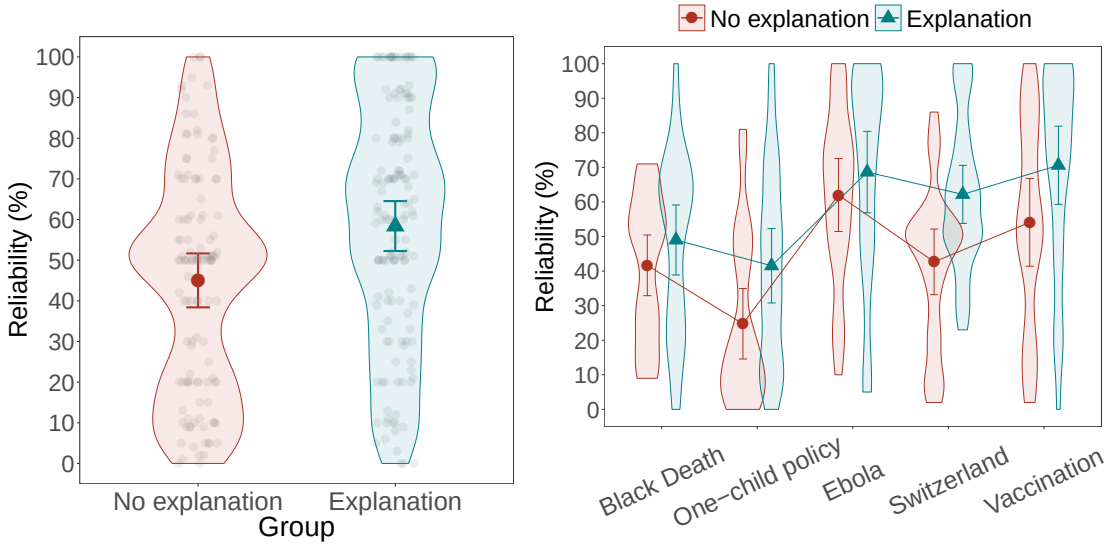
(a) (b)

Figure 4.3: (a) The estimated marginal means (with 95% confidence intervals) from the LMM built on participants' *confidence estimates* in Experiment 7a. Gray points are raw data values (jittered along the x-axis for visibility) with violin plots showing the frequency of the raw data. (b) The observed data means (with 95% confidence intervals) and violin plots for each scenario broken down for each explanation conditions.

4.4.2.2 Reliability

The LMM indicated that reliability estimates in the explanation group ($EMM = 58.39$) were significantly higher than in the no explanation group ($EMM = 45.03$); $t(50) = 2.97$, $p = .005$ (see Figure 4.4a). Further, the inclusion of the predictor for the group in the model led to a significant improvement in model fit ($\chi^2(1) = 8.4$, $p = .004$), compared to just having an intercept as a predictor.

I again found the similar general trend across the five scenarios with some variations in the magnitude (see Figure 4.4b). These variations seem to cor-



(a) (b)

Figure 4.4: (a) The estimated marginal means (with 95% confidence intervals) from the LMM built on participants' *reliability estimates* in Experiment 7a. Gray points are raw data values (jittered along the x-axis for visibility) with violin plots showing the frequency of the raw data. (b) The observed data means (with 95% confidence intervals) and violin plots for each scenario broken down for each explanation conditions.

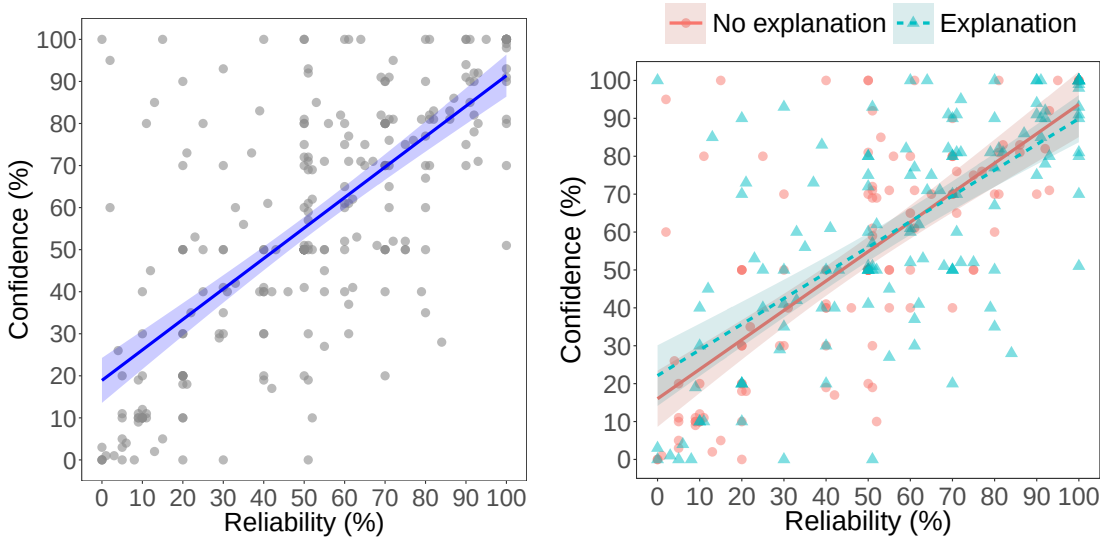
respond to the level of expertise the explainer has. For instance, in the *Ebola* and *the Black Death* scenarios the explainers were a medical practitioner and a member of a research group investigating devastating pandemics in a human history respectively (see Appendix C.2). Plausibly both of these explainers could be considered experts in their fields implying that their reliability is high in the context of these scenarios. In contrast, in *Vaccination* and *Switzerland* scenarios the explainers were students discussing a student project whose reliability in these contexts is arguably low. In the *One-child policy* scenario no information on the explainer's occupational or professional background was provided suggesting no specific level of expertise. However, the scenario con-

text seems to suggest that the explainer and the explainee have only touched upon China's one-child policy in a (casual) conversation suggesting potentially that the explainer is a non-expert. These different levels of expertise (expert vs. non-expert) seem to correspond to the magnitude of the difference between the mean reliability estimates in each explanation condition: in the scenarios where explainer is an expert (high reliability) it seems that the differences in mean reliability estimates between the two explanation conditions are smaller compared to these differences in the scenarios where a non-expert (low reliability) plays a role of an explainer. In Experiment 8 I experimentally manipulate expertise of an explainer to further explore the impact of explanation when explainer's reliability is at different levels.

4.4.2.3 Mediation analysis

A closer look at participants' estimates on the two dependent variables reveals a strong relationship between reliability and confidence in our data (Figure 4.5a): Pearson's correlation $r = .7$, $t(258) = 16$, $p < .001$. One possible explanation of this relationship is that participants simply copied their confidence estimates into their reliability estimates (or vice versa) due to, potentially, their disengagement or misunderstanding of the task. This possibility is explored in Experiment 7b.

Another possibility, however, is that reliability is mediating the effect of explanation on confidence found both in Experiment 6 and Experiment 7a. The initial support for the mediation is readily found in Figure 4.5b where the same strong relationship between reliability and confidence is preserved and unchanged when data is broken down for each explanation condition. I explore



(a) (b)

Figure 4.5: (a) The raw data values of participants' reliability and confidence estimates from Experiment 7a and a linear regression model (with the 95% confidence band). (b) The same data and a linear regression model as in (a) broken down for each explanation condition.

this possibility in more detail here.

Analyses in Sections 4.4.2.1 and 4.4.2.2 suggest that explanation has a significant effect on both confidence and on reliability. Following Baron and Kenny (1986), to explore whether reliability mediates the effect of explanation on confidence I also built a LMM model with both explanation and reliability as predictors of confidence (and same random effects structure as in the above models). If the effect of explanation on confidence in this model was reduced compared to when the only predictor of confidence was explanation (as in Section 4.4.2.1), then this would suggest that reliability is (partially or fully) mediating this effect. I found that when reliability is also included as one of the predictors of confidence, the effect of explanation on confidence disappears

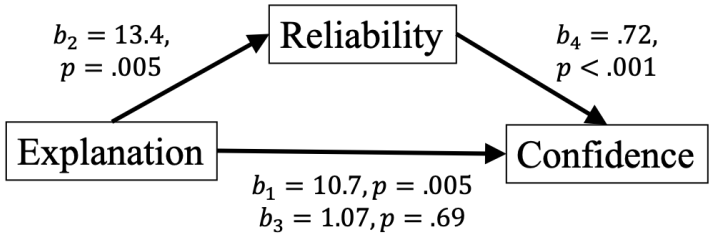


Figure 4.6: Reliability as a mediator between explanation and confidence. b_1, with the related p-value, is the coefficient in a LMM with explanation as a predictor and confidence as a dependent variable (Section 4.4.2.1); b_2 is the coefficient in a LMM with explanation as a predictor and reliability as a dependent variable (Section 4.4.2.2); b_3 and b_4 are coefficients for explanation and reliability respectively in a LMM with explanation and reliability as predictors and confidence as a dependent variable (Section 4.4.2.3). In contrast to b_1, b_3 is minimal and non-significant which suggests that reliability (fully) mediates the effect of explanation on confidence.

($t(257) = 0.39$, $p = .69$) whilst the effect of reliability on confidence is highly significant ($t(257) = 15.4$, $p < .001$) (see Figure 4.6 for a graphical summary of this meditation analysis). Using the "mediation" package in R (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014) I found that the mediation effect is significant ($p = .004$) and that reliability mediates around 90 percent of the association between explanation and confidence. This suggests that a large proportion of the effect that explanation has on confidence is mediated by reliability and that reliability may have a causal effect on confidence. This potential causal effect of reliability on confidence is further explored in Experiment 8 below.

4.5 Experiment 7b

The aim of this experiment was to explore further the possibility which the findings from Experiment 7a suggested: namely, that the strong relationship between participants' confidence and reliability estimates in Experiment 7a was there because participants were simply copying their confidence estimates into their reliability estimates (or vice versa). To that end, instead of eliciting both the confidence and reliability estimates from all participants, I elicited from them either the confidence or the reliability estimates, but not both. If Experiment 7b's results come out to be similar to those in Experiment 7a, then we would more assured that the effect of explanation on confidence and reliability is genuine and that reliability is to some degree mediating the effect of explanation on confidence.

4.5.1 Methods

4.5.1.1 Participants and Design

A total of 121 participants ($N_{\text{FEMALE}} = 81$, two participants identified neither mare nor female, $M_{\text{AGE}} = 34.7$ years) were recruited from Prolific Academic (www.prolific.ac). The selection criteria and the remuneration rate per hour were the same as in the previous experiments. Participants took on average 10 min to complete the experiment.

The design of Experiment 7b is similar to the design of Experiment 7a except that participants in both no explanation and explanation conditions were not asked both the question eliciting their confidence in the claim and the question eliciting their reliability of the explainer but only one of the two questions.

As a result participants were randomly allocated to one of 4 groups: a no explanation group where only confidence rating was elicited ($N = 30$), a no explanation group where only reliability rating was elicited ($N = 30$), an explanation group where only confidence rating was elicited ($N = 30$), and an explanation group where only reliability rating was elicited ($N = 31$).

4.5.1.2 Materials

I used the same scenarios and questions as in Experiment 7a.

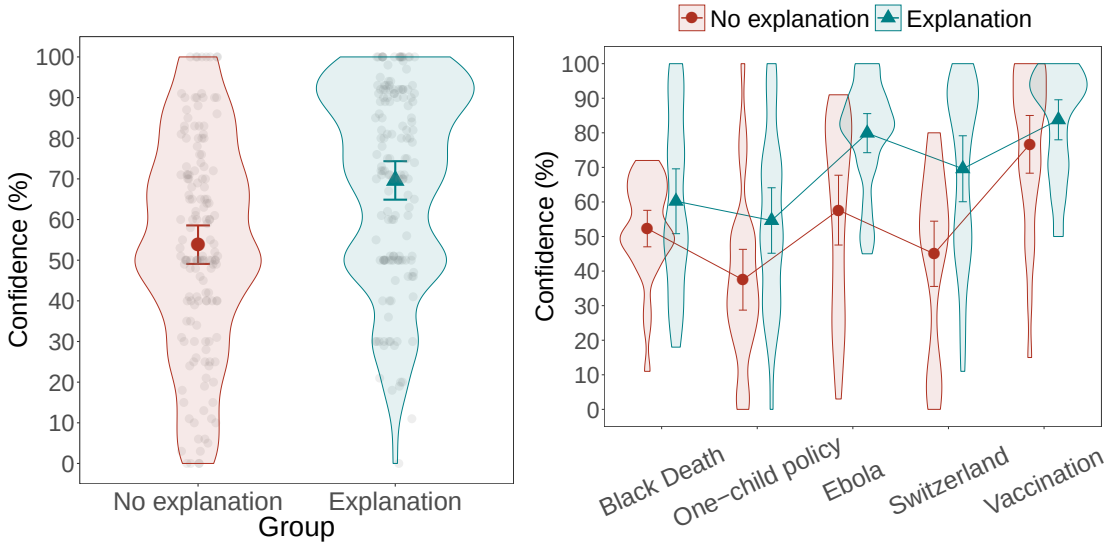
4.5.1.3 Procedure

The procedure was identical to Experiment 7a except that participants were asked only one question rather than two: they were asked either the question about the confidence in the claim or the question about their perceived reliability in the explainer.

4.5.2 Results and Discussion

4.5.2.1 Confidence

The LMM indicated that confidence estimates in the explanation group ($EMM = 69.62$) were significantly higher than in the no explanation group ($EMM = 53.82$); $t(58) = 4.73$, $p < .001$ (see Figure 4.7a). Further, the inclusion of the predictor for the group in the model led to a significant improvement in model fit ($\chi^2(1) = 19.5$, $p < .001$), compared to just having an intercept as a predictor. This trend is also preserved in each of the scenarios (see Figure 4.7b). These results are on par with those from both Experiment 6 and Experiment 7a on confidence and they all follow the same trends.

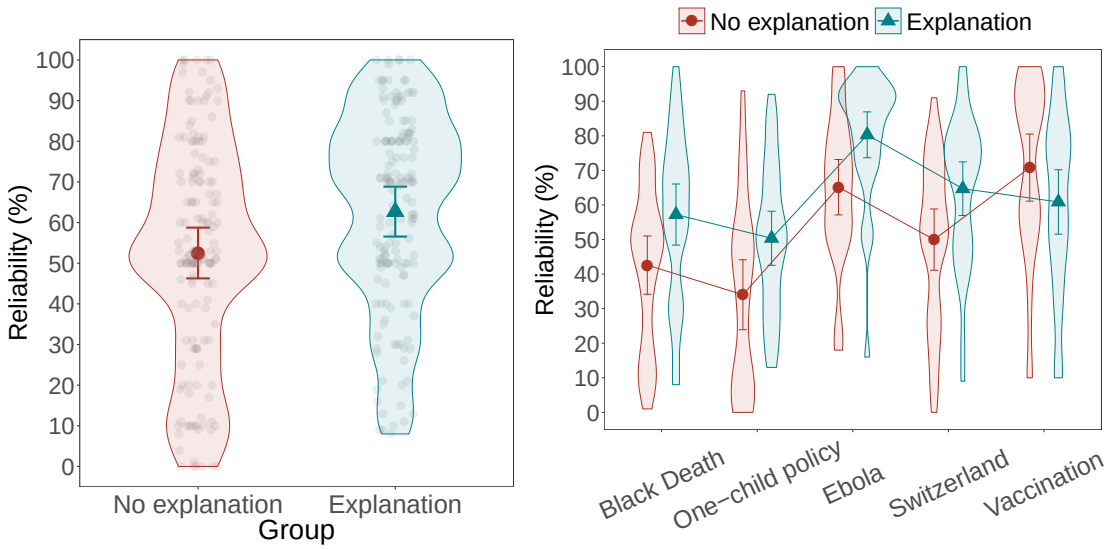


(a) (b)

Figure 4.7: (a) The estimated marginal means (with 95% confidence intervals) from the LMM built on participants' *confidence estimates* in Experiment 7b. Gray points are raw data values (jittered along the x-axis for visibility) with violin plots showing the frequency of the raw data. (b) The observed data means (with 95% confidence intervals) and violin plots for each scenario broken down for each explanation conditions.

4.5.2.2 Reliability

The LMM with reliability as a dependent variable showed that reliability estimates in the explanation group ($EMM = 62.69$) were significantly higher than in the no explanation group ($EMM = 52.51$); $t(59) = 2.33$, $p = .023$ (see Figure 4.8a). Further, the inclusion of the predictor for the group in the model led to a significant improvement in model fit ($\chi^2(1) = 5.36$, $p = .021$), compared to just having an intercept as a predictor. These results also follow the same general trend as those in Experiment 7a, suggesting that participants in Experiment 7a did not simply copy their confidence estimates into their reliability



(a) (b)

Figure 4.8: (a) The estimated marginal means (with 95% confidence intervals) from the LMM built on participants' *reliability estimates* in Experiment 7b. Gray points are raw data values (jittered along the x-axis for visibility) with violin plots showing the frequency of the raw data. (b) The observed data means (with 95% confidence intervals) and violin plots for each scenario broken down for each explanation conditions.

estimates and supporting the idea that the effect of explanation on reliability is genuine.

Zooming in on specific scenarios I found a similar general trend, i.e. participants' reliability estimates were on average higher in the explanation condition than in the no explanation condition, except in the *Vaccination* scenario where the average reliability estimate in the explanation group was lower than in the no explanation group (see Figure 4.8b). Looking into participants' textual explanation of their reasoning for the estimates they provided, I again found hints of the effects of background knowledge. Namely, 11 participants (out of 30) in the no explanation condition of the *Vaccination* scenario wrote that

the claim agreed with their personal (background) knowledge of how vaccines work and all of them provided reliability estimates higher than 60% (this subgroup's average reliability estimate was 86%). Their typical explanations were 'His [explainer's] answer is what I would have said' or 'My understanding [of how vaccines work] is the same as his [explainer's].' The number of participants who provided explanations similar to these and pointed to their background knowledge was only 6 (out of 31) in the explanation condition and all their estimates were also higher than 60% (their average reliability estimate was 84.3%). This shows how (agreement with) people's background beliefs and knowledge can affect their reliability estimates of a person providing an explanation, sometimes even trumping the effects of explanation on reliability.

Together, however, results from Experiments 7a and 7b suggest that reliability is mediating the effects of explanation on confidence, further implying that reliability could also have causal effects on confidence. I explore this the next experiment.

4.6 Experiment 8

Experiment 6 showed that explanations can affect confidence and Experiments 7a and 7b further indicated (i) that explanation also has an effect on reliability and (ii) that explanation's effects on confidence are mediated by reliability, suggesting that reliability could causally affect confidence. In this experiment I explore the this potential causal effects of reliability on confidence. Given the findings in the previous three experiments, I expected that people's confidence estimates will depend on the explainer's level of reliability.

4.6.1 External expertise and perceived expertise

The method that I adopted in this experiment to manipulate reliability was through changing the levels of expertise of the explainer: the higher the level of expertise the higher the reliability. However, in the literature one can find multiple notions of expertise. Thus, before I go on to explore the impact of reliability on confidence, it is worth drawing a distinction between at least two kinds of expertise: an external expertise and a perceived expertise. External expertise is judged by referring to a person's externally measurable criteria: a person's qualifications, their track records of success or their experience of doing a particular activity (see H. Collins & Evans, 2008). For example, doctors are experts according these external criteria as they have required qualifications and potentially relevant experience. This kind of expertise has been found to have significant effect on people beliefs. For instance, the research on the influence of expert testimony on jurors' decision-making suggests that the expert's credentials have a significant effect on jurors decisions (Krauss & Sales, 2001).

Perceived expertise, on the other hand, is not concerned with expert's externally measurable criteria. Rather, it has to do with an expert's general demeanor, such as the internal consistency of their remarks (H. Collins & Evans, 2008). For instance, the judges and jurors would perceive an expert's testimony more believable if it is internally consistent and coherent compared to the one that is less coherent, even though the judges and jurors are not themselves domain exerts. Zemla et al. (2017) similarly point to the distinction between external expertise and perceived expertise and suggest that external expertise may be mediated by a perceived expertise when it comes to the impact of expertise on the goodness of explanations.

In this experiment I manipulated explainers' external expertise. The external expertise, I believe, would have an effect on participant's perceived expertise of the explainers, which is what is being measured by asking participants to provide their estimates of the reliability of the explainers in the scenarios (see also Zemla et al., 2017). Experiments 7a and 7b suggested that the presence/absence of an explanation for a claim has an impact on the explainer's perceived reliability. I thus expected that the impact of the external expertise will be attenuated by explanation resulting in a lesser effect of the reliability on the confidence in claims in the conditions where the explanation for the claim was provided.

4.6.2 Methods

4.6.2.1 Participants and Design

A total of 161 participants ($N_{\text{FEMALE}} = 112$, one participant identified as neither male nor female, $M_{\text{AGE}} = 36.5$ years) were recruited from Prolific Academic (www.prolific.ac). The selection criteria and the remuneration rate per hour were these same as in the previous experiments. Participants took on average 14.8 min to complete the experiment.

A between-participant design was adopted and participants were randomly allocated in one of 2 (no explanation or explanation) $\times$ 2 (reliability: low or high) = 4 groups ($N_{\text{NO\_EXPL\_LOW}} = 40$, $N_{\text{NO\_EXPL\_HIGH}} = 42$, $N_{\text{EXPL\_LOW}} = 40$, $N_{\text{EXPL\_HIGH}} = 39$). All participants were asked to provide estimates on two dependent variables: confidence and reliability.

4.6.2.2 Materials

I used the same 5 scenarios as before with some further modifications so that the explainer's reliability is either high or low. This was done in a way that in the preamble of each scenario the explainer was introduced either as a domain expert (high reliability) or a novice/lay person (low reliability). For example, *the Black Death* read as follows (note that the text in parentheses appeared only in the explanation condition):

> *Preamble in the low reliability condition*: Dave and Jimmy are high school students who are assigned a student project to find out as much as they can on one of the most devastating pandemics in human history, namely the Black Death.
>
> *Preamble in the high reliability condition*: Dave and Jimmy are senior researchers at a well-established institute for global health and part of the project investigating devastating pandemics in human history. During a planning meeting they touched upon the Black Death.
>
> *The rest of the scenario was the same for both the low and high reliability conditions.*
>
> **Dave**: Millions of people died from the Black Death in the 14th century. I think our project should in part focus on how the Black Death ended.
>
> **Jimmy**: Yes, I agree. Do you already have an idea regarding how the Black Death came to an end?

Dave: The Black Death subsided mostly through the use of quarantines.

(**Jimmy**: How so?

Dave: People mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources would traditionally escape to the country, far away from the Black Death-infested cities.)

The expert explainers in other scenarios were: an immunologist (*Vaccination* scenario), an experienced policy-maker who specialized on East Asia (*One-child policy* scenario), a medical practitioner who was a part of the Doctors Without Border team in West Africa treating various epidemic diseases (*Ebola* scenario), a history professors who have been awarded a research grant for a project on armed neutrality in World War II (*Switzerland in WWII* scenario). The non-expert explainers were: a subway operator (*Vaccination* scenario), a person who has just started their undergraduate studies in philosophy (*One-child policy* scenario), a non-medically educated person who read in the news about a team of doctors in West Africa who contracted Ebola (*Ebola* scenario), a high school student (*Switzerland in WWII* scenario). For full materials see Appendix C.3.

4.6.2.3 Procedure

The procedure was exactly the same as that of Experiment 7a with each participant being asked both the confidence question and the reliability question. The main reason for including the reliability question in addition to the confidence question was to check the success of the reliability manipulation: participants'

reliability estimates would indicate if they have accepted high/low reliability of the explainer that I aimed to communicate in the scenarios.

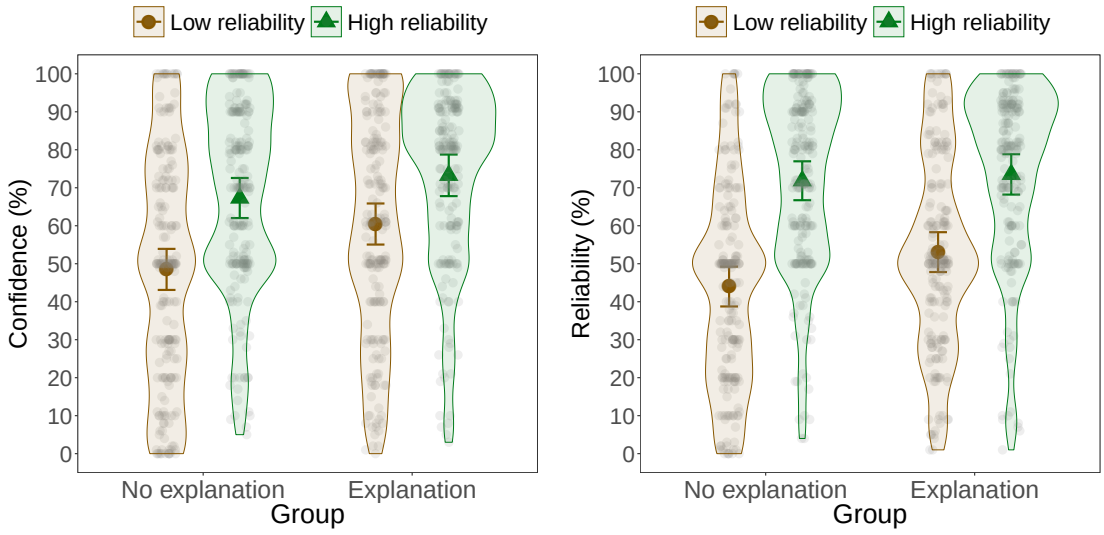
4.6.3 Results and Discussion

A separate analysis was conducted for each dependent variable.

4.6.3.1 Confidence

To test the effect of explanation and reliability on people's confidence in statements, I built an LMM with explanation and reliability as fixed effects and a random intercept for each participant. I found a main effect of explanation ($t(157) = 3.28$, $p = .001$) and of reliability ($t(157) = 5.79$, $p < .001$), and no interaction between the fixed effects ($t(157) = -1.1$, $p = .28$). This suggests that both explanation and reliability have (additive) causal effects on confidence.

Experiment 7a suggested that the effects of explanation were of lesser magnitude when the reliability of the explainer was high. This was explored more directly here. For the low reliability conditions I found that, the explanation group's confidence estimates ($EMM = 60.5$) were significantly higher than the no explanation group's ones ($EMM = 48.5$), $t(78) = 2.74$, $p = .008$. This was not the case, however, in the high reliability conditions: there the explanation group's confidence estimates ($EMM = 73.3$) were not significantly higher than the no explanation group's ones ($EMM = 67.3$), $t(79) = 1.81$, $p = .075$ (see Figure 4.9a). These results suggest that the effects of explanation on confidence are large when the reliability of the explainer is low, but that they disappear when the reliability of the explainer is high. We can see this effect in Figure 4.10 where as the difference in mean estimates between the two explanation conditions is

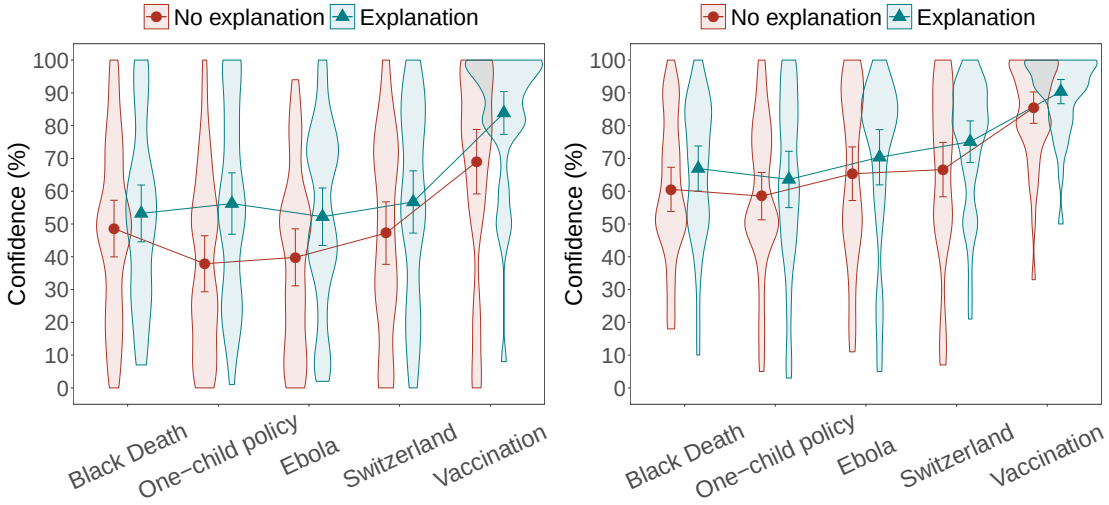


(a) (b)

Figure 4.9: (a) The estimated marginal means (with 95% confidence intervals) from the LMM built on participants' *confidence estimates* in Experiment 8. (b) The EMMs (with 95% confidence intervals) from the LMM built on participants' *reliability estimates* in Experiment 8.

smaller in the high reliability conditions for each cover story.

Post-hoc contrasts on the differences between the four group's mean confidence estimates provided us with a more detailed view. Specifically, I found that there was no significant difference between the means of the explanation and low reliability group and the no explanation and high reliability group (see Figure 4.12a). This finding points at the capacity of adding an explanation and increasing reliability to increase the confidence. Namely, this result suggests that a low reliability (non-expert) explainer who provides a (good) explanation to their claim would have the same impact on explainee's confidence as if that claim was provided by a high reliability (expert) explainer who did not provide any further explanation for their claim. Thus, increase in reliability and



(a) (b)

Figure 4.10: (a) The observed data means (with 95% confidence intervals) for participants' *confidence estimates* in each scenario in *the low reliability* condition. (b) The observed data means (with 95% confidence intervals) for participants' *confidence estimates* in each scenario in *the high reliability* condition.

providing an explanation may have similar (boosting) effect on confidence.

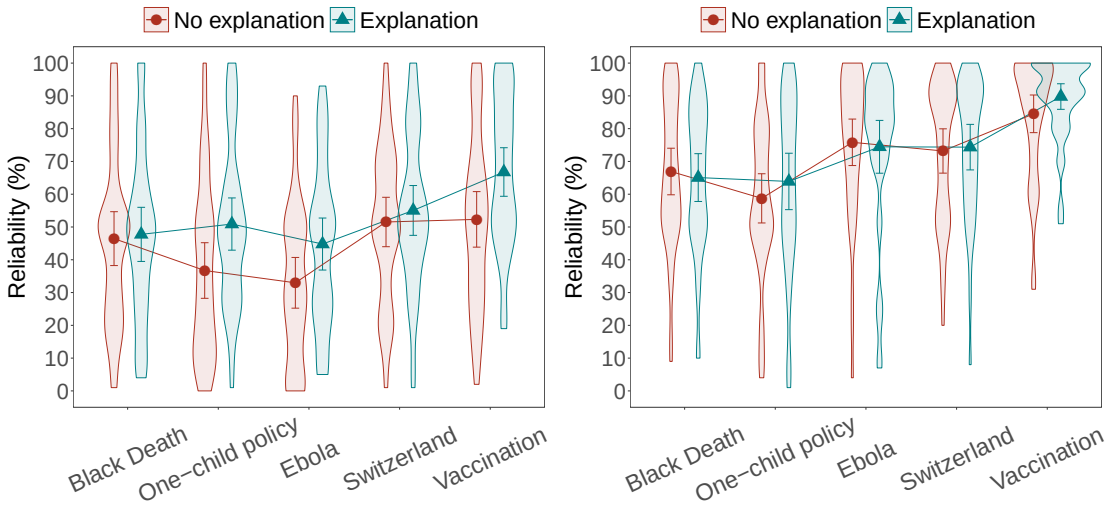
4.6.3.2 Reliability

To test the effect of explanation and reliability on people's reliability estimates, I build an LMM with explanation and reliability as fixed effects and a random intercept for each participant. I found a main overall effect of explanation ($t(157) = 2.03$, $p = .045$) and of reliability ($t(157) = 9.12$, $p < .001$), and no interaction between the fixed effects $t(157) = -1.39$, $p = .17$). Further, the inclusion of the predictors for the model led to a significant improvement in model fit ($\chi^2(3) = 72.3$, $p < .001$), compared to just having an intercept as a predictor. Similarly to the above findings on confidence there was an effect

of explanation and a highly significant effect of the reliability manipulation on participants' reliability estimates suggesting that participants in the high reliability condition provided higher estimates regarding the explainer's reliability compared to those in the low reliability condition, as expected.

I further found that in the low reliability conditions, the explanation group's reliability estimates ($EMM = 53.1$) were significantly higher than no explanation group's estimates ($EMM = 44$), $t(78) = 2.16$, $p = .034$; and in the high reliability conditions, the explanation group's reliability estimates ($EMM = 73.5$) were not significantly higher than no explanation group's estimates ($EMM = 71.8$), $t(79) = 0.52$, $p = .61$ (see Figure 4.9b). This suggests that when the reliability of the explainer is low, then that reliability could be increased if the explainer also provided an explanation. However, if the explainer is already highly reliable (an expert), then explainer additionally providing an explanation will not significantly increase their reliability. This general effect was also reflected when focusing of specific scenarios as well (see Figure 4.11).

Contrasts (see Figure 4.12b) show that the main driver of participants' reliability estimates was whether they were in the low or high reliability conditions: in the low reliability condition an explainer providing an explanation did, as expected, seem to result in participants providing higher reliability estimates than if there was no such explanation. However, unlike in the case of participants' confidence estimates, providing an explanation in a low reliability condition did not lead to similar reliability estimates in the high reliable condition where no explanation of their claim was provided. Instead, these reliability estimates were fully driven by whether participants were in the low or in the high reliability condition.

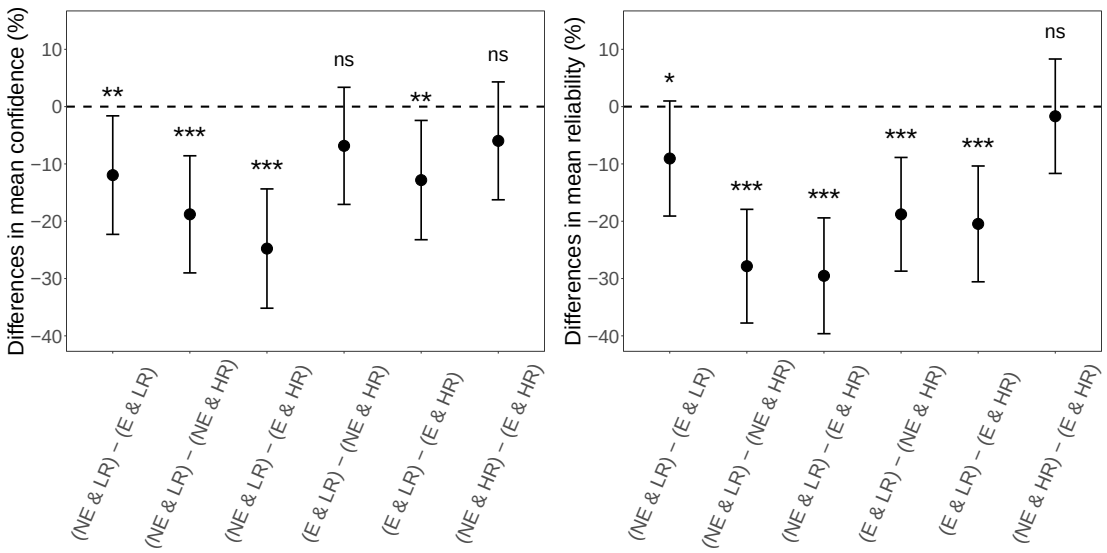


(a) (b)

Figure 4.11: (a) The observed data means (with 95% confidence intervals) for participants' *reliability estimates* in each scenario in *the low reliability* condition. (b) The observed data means (with 95% confidence intervals) for participants' *reliability estimates* in each scenario in *the high reliability* condition.

4.7 General discussion

I carried out four experiments where I aimed to explore the relationship between explanation understood as links between claims and data, reliability, and confidence. Experiment 6 provided evidence that explanations (of inferences) do have a impact on our confidence; more specifically, providing an explanation increases confidence in a claim. This is in line with the previous literature on explanations of evidence and show that the same holds for the explanations of reasoning processes. Experiments 7a and 7b introduced an aspects of explanation, namely reliability, that becomes apparent only when explanation is fully considered as a social act. These two experiments tested the impact of



(a) (b)

Figure 4.12: (a) The contrasts for the different combinations of the explanation and reliability conditions for participants' mean *confidence* estimates. (b) The contrasts for the participants' mean *reliability* estimates. NE: no explanation; E: explanation; LR: low reliability; HR: high reliability. *P*-value indicators: ns := $p > .05$, $* := p \leq .05$, $** := p \leq .01$, $*** := p \leq .001$. All *p*-values were corrected for multiple comparisons using Benjamini and Hochberg's false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995).

providing an explanation on both confidence and reliability. The findings from these experiment provide evidence that a (good) explanation increases both the confidence in a claim (in line with Experiment 6) as well as the perceived reliability of an explainer. Furthermore, these two experiments indicated that much of the effect of providing an explanation on confidence is mediated by the reliability, suggesting that the reliability of an explainer may have a causal effect on confidence in a claim. Experiment 8 explored this potential causal effect of reliability on confidence. The findings from Experiment 8 suggested that (i)

the effect of providing an (good) explanation on confidence is larger when the reliability of an explainer is low than high, (ii) a non-expert providing a (good) explanation may have very similar effect on confidence as an expert who did not provide any further explanation for their claim, (iii) participants' reliability estimates were mostly guided by the level of expertise of an explainer, with explanation only having an impact when the expertise was low, and (iv) in contrast to the findings regarding the effects of reliability and explanation on the confidence estimates, a non-expert providing a (good) explanation did not lead to a similar effect on the reliability estimates as an expert who did not provide an explanation for their claim; rather, the reliability estimates in this case were driven purely by the levels of expertise.

These findings complement and further add to the previous research on explanation. For example, Koehler (1991) conducted experiments where participants were asked to provide explanations of hypotheses and found that explanation boosts confidence. In the above four experiment, I closely replicated these findings with a different paradigm where participants were given explicit explanations without being asked to produce them. Also, these findings on the effects of explanations of reasoning processes understood as links between claims and data replicated those from the previous studies on IBE and explanatory virtues where focus was on the notion of explanations of evidence (outcome) (e.g. Douven & Schupbach, 2015; Lombrozo, 2007, 2016; Lagnado, 1994). Additionally, these findings extend onto everyday explanations of real-world events that, arguably, carry more ecological validity. The work on everyday explanations is still very much in its infancy and often limited to correlations analyses of different explanatory criteria (Zemla et al., 2017). All four

experiments included a randomized allocation of the participants to different conditions and went beyond the correlations analyses thus further contributing to our understanding of the impacts of everyday explanations on our beliefs.

This was also one of the first studies to empirically investigate some of the social aspects of explanation, in particular the reliability of an explainer. It showed the interplay between providing an explanation and the reliability of an explainer, specifically in the context of everyday explanations. Apart from the expected results that providing a (good) explanation that is coming from an explainer with a high reliability, the findings suggest that in certain cases a non-expert providing a (good) explanation may have a very similar impact on our beliefs as an expert just claiming something true without providing an explanation for the claim. These results seem to align with the predictions of some of the formal models of source reliability, such as the Bovens and Hartmann (BH) model (Bovens & Hartmann, 2003). According to the BH model, the higher the the reliability of a source the larger the impact of evidence on the confidence in a hypothesis. The findings from Experiment 8 suggested that this is the case in both the no explanation and the explanation condition. Furthermore, if we slightly modify the BH model to include explanation as an additional variable whose effect is (partially) mediated by the reliability of the explainer, then one would expect that in such a model the impact of explanation is higher when the reliability is low compared to when the reliability is high, and given a certain plausible parametarizations of that model one could expect that the impact of explanations when the reliability is low is similar to the impact of the high reliability source that is missing an explanation. These are, however, only (plausible) conjectures and call for further exploration.

The effects of reliability on people's beliefs although under-explored could potentially extend to and account for some of the finding regarding the effects of explanations in the literature. For example, Weisberg, Keil, Goodstein, Rawson, and Gray (2008) showed that added irrelevant neuroscience information had a particularly striking effect on non-experts' judgments of bad explanations. Namely, non-experts judged these explanations significantly more satisfying than the bad explanations without irrelevant the neuroscience information.[2] Furthermore, the Experiment 2 results from Weisberg et al. (2008) suggested that there was no difference between the satisfaction ratings of the good explanations without any irrelevant neuroscience information and the bad explanations with this information. Weisberg, Taylor, and Hopkins (2015) provide a few potential explanations of why this may be. However, another explanation that they have not considered comes from the findings of the four experiments presented in this chapter. Namely, if, as it seems plausible, the presence of the irrelevant neuroscience information indicated to the participants that the explanation is coming from a reputable/reliable source with potentially expertise in neuroscience, then, in line with the findings from Experiment 8, one could expect that the participants' rating of the bad explanations would be higher when such information is present and even potentially having similar ratings as the good explanations that lack such information.

This also suggests that there are different ways to communicate to the participants the reliability level of the explainer or the source of explanation. In

[2]Similar results were found when non-experts were asked to judge the quality of research based on the abstract of papers. Eriksson (2012) found that the non-experts judged the research to be of higher quality if it included equations, even though the equations that did not make sense in the context of research.

the experiments presented in this chapter I have done that by communicating to the participants different levels of the external expertise. The results from Experiment 8 suggest that this was successful in manipulating the participants' perceived reliability or expertise of the explainers. Another potential way of manipulating the perceived reliability could be through including technical information relevant to the domain as suggested above. However, neither of these two ways manipulate reliability directly: they both manipulate some other factor (external expertise or the presence of technical information) in order to affect the (perceived) reliability of the explainer. Hahn et al. (2009) suggest a potentially more direct way to manipulate reliability. Namely, they manipulated the reliability of the source by communicating to the participants that the information came from a reliable source (e.g. a respected journal Science) or from an unreliable source (e.g. a Facebook post from the Gossip Mill Mzansi Facebook group). These and other ways to manipulate the reliability of the source should be investigated in the future research on explanations.

Throughout this chapter we have seen hints of the effects of the prior or background knowledge on both the confidence in claims and reliability of an explainer. We have seen that in some scenarios (e.g. *Vaccination* and *Ebola* scenarios) the participants' confidence and reliability estimates in both the no explanation and explanation conditions were high and that the different in their estimates between the conditions was small. The participants' textual responses show that they provided high estimates to the confidence and the reliability questions in these scenarios as the claim and explanations were in line with their prior knowledge. This suggest that the effects of explanation and reliability may be attenuated if the effect of the prior knowledge is high.

This is in line with the previous literature. Koehler (1991) and C. A. Anderson and Sechler (1986) suggest that people who already have a strong opinion about a topic are unlikely to be influenced by an explanation. However, people who do not have a strong opinion or do not have a formed impression about a topic will especially be subject to the effects of explanation. In particular, the effects of explanation are lower if people were familiar about the topic in question. This could be because people are less likely to seek an explanation for something that is already familiar to them (Lombrozo, 2012) and/or because people already believe in the claim with a high degree of confidence (Thagard, 1989). Similar effect of the prior knowledge on explanation have been observed in category learning (Williams & Lombrozo, 2013). The results from Jarvstad and Hahn (2011, Experiment 2) suggested that the participants' prior confidence in the claims affected the source reliability judgements, with a source being judged more reliable if it provided a statement in which participants had higher prior confidence.

4.8 Conclusions

In this chapter I have explored the effect of one of the aspects of explanations being communicative acts, namely the reliability of an explainer. Specifically, I have explored the interplay between explanation understood as links between claims and evidence and the reliability of an explainer as well as their effects on confidence in claims. I have found that, in line with the literature, providing an explanation affects our confidence, but that effect may be mediated by reliability. Furthermore, I have found that the effects of an explanation are higher in the low reliability contexts, and that the effect of explanation in the low reliabil-

ity setting can be as high as the impact of a (higher) reliability of an explainer in contexts where an explanation was not provided.

This is only one study on the effects of reliability in the context of explanations understood as links between claims and evidence has been under-explored and further studies should be conducted. Nonetheless, it provides fruitful ground for further exploration. The reliability of an explainer, however, is only one of the aspects of explanation when considered as communicative acts. Other aspects like 'normality' or the effects of explanation in learning are also understudied, especially in the context of explanations of inference processes understood as links between claims and evidence. These point to further research avenues that could be explored.

5

General discussion

In this thesis I have explored the relationship between argument and three notions of explanation. This exploration led into discussions regarding the strategies one might adopt in causal-probabilistic reasoning, probability interpretations, extending the algebra, goodness of explanations, and reliability. Specifically, I examined in more detail the interplay between argument and explanation as highlighted by the most fundamental argument scheme in the literature, the so-called argument from sign. Here, 'critical questions' about

potential alternative explanations for the 'sign' (evidence) that are essential to the scheme-based tradition's normative guidance can be recast in a normative Bayesian framework as the phenomenon of 'explaining away'. This led me to a detailed investigation of strategies lay reasoners use for 'explaining away'. In particular, I examined strategies adopted in causal probabilistic reasoning and their relationship with different probability interpretations. This closer look at explaining away then led on to a fundamental, but typically overlooked, normative and descriptive 'gap' not just for explanation but also for argument, namely what happens when new variables emerge, that is, the algebra must be extended. Finally, I explored goodness of explanations and the impact of explanations on perceived reliability.

In this chapter I restate the main findings and provide further implications and future research directions suggested by the findings.

5.1 Brief overview of experimental data

5.1.1 Argument and explanation in causal reasoning

To explore the relationship between explanations and arguments in explaining away, I carried out three experiments utilising a novel methodology. The findings suggested that participants understood the explaining away structure and accepted the parameters communicated to them. Despite that, the main findings echoed those of the extant literature as participants systematically violated the normative account of explaining away. Specifically, in Experiments 1 and 2 pitfalls were most evident in diagnostic reasoning and direct explaining away.

Further, findings from all three experiments suggested that deviations from

the normative model 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. All three experiments found that a significant proportion of participants split the probability space among the causes in line with diagnostic split reasoning. This was most evident when participants reasoned with low priors. The findings from all three experiments also suggested that a large number of participants remained at the priors when answering diagnostic reasoning and direct explaining away questions, with the proportions of participants who did not change their probability estimates varying as predicted by the propensity hypothesis.

In the second part of Chapter 2 I explored people's reasoning in structures that extend explaining away. Specifically, I explored people's sequential diagnostic reasoning in an explaining away structure with two effects. This structure also allowed for the possibility of learning the algebra in a sequential way. The findings from Experiments 4 and 5 suggested that people are sensitive to the different diagnosticities of the evidence in this extended explaining away structure, both when the probabilistic information regarding diagnosticity was communicated through numerical probability estimates and through verbal probability expressions. However, I found that people do not follow the predictions of the normative (full) CBN model and sometimes provided estimates that went against even the qualitative predictions of the normative model. This suggests that although people were sensitive to the different diagnosticities of the evidence, it is perhaps not these diagnosticities that led them to provide specific estimates for the two causes. Additionally, results suggested that there was no difference between the groups who were presented the full algebra

from the beginning and the groups who were presented the algebra in a sequential way and that there was little support for either of the two modeling strategies considered, i.e. the split and the full model.

The propensity interpretation and the diagnostic split were also explored in this part of Chapter 2. The results suggested that the diagnostic split hypothesis accounted for a large proportion of the responses in Experiment 4, although some of these responses could have been due to participants' providing .50 as their estimate to communicate that they do not know the answer. In Experiment 5, the diagnostic split hypothesis account for a significantly smaller proportion of the estimates. Overall, the propensity interpretation accounted for a significantly smaller proportion of participants estimates in Experiments 4 and 5 compared to Experiments 1–3.

5.1.2 Explaining the argument

In Chapter 3 I discussed the theoretical background of explanations. In particular I distinguished between two notions of explanations with a focus on causal Bayesian networks (CBNs) and reviewed the literature on explanatory virtues. One of the upshots of the theoretical discussion was that it is not clear how certain aspects of explanations, such as explanatory virtues, map onto explaining inferences in CBNs.

I thus conducted a case study where I explored how a human (expert) would explain inference in CBNs. I found that experts are capable of explaining inferences processes in CBNs and that their explanations often have the function of a link or a bridge between the target node (claim/hypothesis) and the evidence. Further, multiple features of the philosophical, psychological,

and cognitive science literature are reflected in these explanations: a focus on causal explanation for a probabilistic system; the directional nature of explanation (its asymmetry); indications of pragmatic sensitivity in that hypotheticals are used to express relevant 'contrasts'; and, an emerging notion of the explanatory virtue simplicity in the use of the Markov blanket.

5.1.3 Social explanation

In Chapter 4 I introduced explanations as communicative acts and explored the effects of one of the aspects of explanations understood as such, namely the reliability of an explainer. I carried out four experiments where I aimed to explore the relationship between explanation, reliability, and confidence. The findings from the four experiments suggested that (i) (good) explanations (of inferences) do have an impact on our confidence, that is, providing an explanation increases confidence in a claim; (ii) a (good) explanation increases the perceived reliability of an explainer; (iii) the effect of providing an explanation on confidence is mediated by that reliability; (iv) the effects of providing a (good) explanation on confidence is larger when the reliability of an explainer is low than high; (v) a (good) explanation that was provided by a non-expert may have very similar effects on confidence as a (high) reliability of an expert who did not provide any further explanation for their claim; (vi) people's reliability estimates were mostly guided by the level of expertise of an explainer, with explanation only having a impact when the expertise was low; and (vii) a non-expert providing a (good) explanation did not lead to a similar effect on the reliability estimates as an expert who did not provide an explanation for their claim; rather, the reliability estimates in this case were driven purely by

the levels of expertise.

5.2 Implications and future directions

5.2.1 Diagnostic split, propensity interpretation, and other probability interpretations

The results from Chapter 2 showed that in some situations providing an explanation of evidence will have a limited effect on the strength of an argument, i.e. the posterior probability of a claim after additionally learning that there is a competing explanation. Specifically, the majority of people will just stay at the prior probability of the claim or split the probability space between the claim and explanation. This, however, does not imply that people's estimates are always going against the normative framework. We have seen that if people are providing estimates that are in line the propensity interpretation when the priors are above .5 and in line with the diagnostic split reasoning otherwise, then their estimates will be within .1 of the normative answer about 2/3 of the time. One could, thus, argue that even though the two strategies are incompatible with the normative predictions, people who reason in line with the two strategies are approximately rational most of the time.

In non-deterministic set-ups, however, the effects of the propensity hypothesis and the diagnostic split seem to be significantly reduced. These are set-ups 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. Nonetheless, Chapter 2 presented one of the few studies

on the potential effects of probability interpretations. It suggested that in certain argumentation contexts the effects of explanations on the strength of the argument depend on how people interpret probabilities.

Further research should consider other argument schemes and the effects of the propensity interpretation and/or diagnostic split reasoning in these contexts. I have suggested that even in very simple contexts where participants were assessing the impacts of evidence on only one claim, the effects of the propensity interpretation may be significant especially when the set-up is deterministic, the prior probabilities are well-established, and there are clear causal-mechanistic relations between the claim and evidence. The propensity interpretation, however, is not the only way on can understand probabilities (Hájek, 2012). For example, the frequency interpretation is another interpretation that is prevalent in both every day contexts (e.g. rolling an even number on a die) and in science (e.g. traditional/frequentist statistics). Many studies have employed a frequency format response scale to elicit probability estimates from participants (e.g. Gigerenzer & Hoffrage, 1995), which may have emphasised the frequency interpretation more and thus lead participants to provide estimates whilst considering probabilities as frequencies. Nonetheless, studies directly assessing the effects of probability interpretations (including the frequency interpretation) are still very few.

5.2.2 Extending algebra

Both theoretical and empirical research on extending the algebra is limited, not just for explanation but also for argumentation. The empirical results from Chapter 2 suggested that people do not differentiate between learning that a

certain event that we knew was possible took place and learning about the possible existence of an event and including it into our belief system. However, these are only first studies to explore potential effects of extending algebra. Further research is warranted.

On the theoretical side, I have explored two possible ways to model extending algebra. The Bayesian framework, however, does not provide suggestions or put constraints on how models should incorporate new events. Suggestions on how to address the problem may come from some of the distance-based approaches (e.g. Eva et al., 2019). It is important to note, however, that we cannot be aware of all possible events that could potentially exist. Thus, any approach tackling this issues will have to be flexible enough to allow for extending algebra to add new, yet unaccounted for events.

5.2.3 Explanations, arguments, and AI

In Chapter 3 I presented and explored a different notion of explanation to that explored in Chapter 2; namely, the notion of explanations of reasoning/inference processes. The literature review suggested that this notion is mostly found in the computer science, but that it is also pertinent to psychology as it relates to explaining the argument, provided that argument can be represented via CBNs. I have then devised a case study that provided us with an initial sense of the kinds of explanations a human (expert) may produce and, potentially, prefer in the context of explanations of reasoning in CBNs, and more generally in AI systems. More specifically, the case study makes a start at bridging between computer science, and the philosophical and psychological literature on what counts as 'good reasoning' by eliciting explanations

from human experts and it provides concrete direction for future algorithm construction. The study also pointed out that while the literature from philosophy and psychology is a helpful starting point, more context specific work would be required to address the question of what would be the most appropriate explanation of the reasoning processes in a particular AI domain.

The discussion on the three dimensions of explanation showed that explanation also have a strong social component. In Chapter 4 I have explored one aspect of that component, namely the reliability of the source of explanation. The results showed the interplay between providing an explanation and the reliability of an explainer, specifically in the context of everyday explanations. I also pointed out that the Bovens and Hartmann (BH) model (Bovens & Hartmann, 2003) may be able to capture at least some aspects of this interplay as well as the ways these findings can potentially account for some of the prior findings regarding irrelevant information in explanations (Weisberg et al., 2008, 2015).

The findings from Chapter 4 have also pointed to the importance of background knowledge or prior beliefs in assessing the effects of explanations on confidence or the strength of arguments. Namely, people who found that claims and explanations aligned with their prior beliefs deemed the effects of providing an explanation on their confidence in these claims small. This is in line with the previous literature in psychology (C. A. Anderson & Sechler, 1986; Jarvstad & Hahn, 2011; Lombrozo, 2012; Koehler, 1991; Thagard, 1989; Williams & Lombrozo, 2013)

The effects of background knowledge on explanation have also been pointed out in AI literature as well (Lacave & Díez, 2002; Miller, 2019). Dif-

ferent users of an AI system can have different knowledge levels about the domain that is modeled and/or different knowledge levels about the models themselves. Thus, different users would require different kinds of explanations that would reflect the differences in the background knowledge. If an AI system is to produce a convincing explanation, it would need to 'know its audience' and tailor explanations accordingly. Thus, studying the effects in a more direct way should be pursued in future research.

5.2.4 Explanations and trust

The findings regarding explanations and reliability from Chapter 4 can also inform the research on (machine-produced) explanations in AI. Because explanations and source reliability jointly impact of our beliefs, these interactions are likely to be consequential for the extent to which the conclusion of an AI system being explained is itself perceived to be true. The literature in AI, in particular recommender systems, has long recognized relationship between trust and explanation (Zhang & Chen, 2020). The majority of research suggests that providing an explanation improves user's trust in an AI system (Herlocker et al., 2000; Sinha & Swearingen, 2002; Symeonidis, Nanopoulos, & Manolopoulos, 2009). However, the situation seems more intricate as more transparent systems do not always lead to increase in trust (Cramer et al., 2008), and sometimes poor explanations can lead to reduced acceptance of the AI systems (Herlocker et al., 2000). To explore the interactions between explanations and trust, in addition to manipulation transparency of AI systems, one would also need to experimentally manipulate the level of trust users have in them.

The results from Chapter 4 seem to naturally translate into the AI context.

For example, consider the following findings: (i) providing an explanation for a claim increases not just people's confidence in the claim but also the perceived reliability of the person providing an explanation as compared to when there is no such explanation and (ii) providing an explanation has a significantly greater impact on the confidence and reliability when people's initial (prior) reliability of the source is low compared to when that reliability is high. In the context of AI, these results suggest that providing a (good) explanation of an AI system's decisions will arguably increase people's confidence in/acceptance of these decisions as well as people's perceived the reliability/trust of the system. In particular, the impact of providing an explanation will be greater (and most useful) if people's initial perceived reliability/trust of an AI system is low. These claims, however, should be empirically explored in the context of AI.

5.2.5 Trust and fidelity

A recent surge of interest in explanations of black-box deep learning models has significantly pushed the horizons of explainable AI, but at the same time it has also introduced the *fidelity* problem. Namely, unlike explanations of CBN where original CBN models could be used to generate explanations (either as justification of evidence or as explanation of reasoning processes), deep learning models are not transparent enough for either a lay or an expert human user to be able to explain the models' outputs; rather, one resorts to explanation models that are independent of the deep learning models to generate explanations of these black-box models' decisions after these decisions have been made, i.e. post-hoc (Ribeiro, Singh, & Guestrin, 2016; Zhang & Chen, 2020). The explanation models are often model agnostic as they should be able to ex-

plain decisions of any (black-box) model. Post-hoc model agnostic explanation models have certainly furthered the work on explanation in AI, but they have also prompted questions regarding the degree to which the explanations generated by models reflect the real mechanisms that generated decisions of a deep learning model: i.e. they have raised questions regarding the fidelity of explanation models (Sørmo et al., 2005; Ribeiro et al., 2016). In the literature, the trade-off between fidelity and interpretability of explanation models is often acknowledged: the higher the fidelity of an explanation model to the black-box model the lower the interpretability of that model and its transparency to a human user (Ribeiro et al., 2016). This however brings trust and reliability into the consideration. On the one hand, if higher interpretability is to increase trust and reliability, then trust may be negatively affected by higher fidelity. On the other hand, if users expect higher fidelity explanation models, then lower fidelity may now negatively affect trust. This potentially interesting relationship between fidelity and trust is another open issue related to the interplay between a user and the system that should be addressed in the future research on explanations.

5.2.6 Further research avenues

The effect of the communication on the reliability of the source/trust is not the only way in which explanations, when considered as communicative acts, alter beliefs about what it is that is being explained. For example, does providing an explanation constrain and/or make less ambiguous the underlying (causal) structure of the world that the explainee had in mind before receiving the explanation (or in the case of a CBN, does providing an explanation restrict

the number of potential CBN structures that the explainee has mind) (c.f. Bes, Sloman, Lucas, & Raufaste, 2012)? How does providing an explanation of an (ab)normal event in a causal chain of events reflect on our perceptions of that explanation (Kirfel, Icard, & Gerstenberg, 2020)? Does a detailed explanation of a usual and obvious succession of events make that explanation less preferred or worse compared to a less detailed explanation (Bechlivanidis et al., 2017)? All these questions call for further investigation and can have implications for the explainable AI project.

The three dimensions of explanation have carved the explanation space into 8 intersection points. Out of these 8 points I have explored only three of them, and only one aspect of understanding explanations as communicative acts. In this thesis I have not explored explanations as processes. The research on explanations as processes is itself limited and it calls for further exploration, specifically in social contexts where the social interaction may facilitate the cognitive activity of explaining something. Similarly, considering the relationship between arguments and explanations as processes would be particularly interesting because argumentation as an activity or process includes a dynamic between the interlocutors. It would be worthy of exploration to consider the effects of explanations as processes on arguments, particularly when one is fully engaged with the cognitive activity of producing an explanation, but failing to actually produce one. I also have not explored explanations of outcomes that are products and that are made in a social context. Although potentially interesting, exploring this intersection of the explanation space may prove to be more difficult. This is because in social contexts it may be more challenging to just provide explanations of an outcome without going into and explaining the

inference processes that lead to that outcome.

5.3 Conclusions

In this thesis I have explored the relationships between arguments and explanations. In doing so I have empirically explored the effects of probability interpretations and extending algebra on the confidence or the strength of arguments. I have reviewed the theoretical background on explanations from different fields, including philosophy, psychology and computer science, and summarized some of the findings from this review via the three dimensions of explanation. I have also explored the effects of some of the aspects of understanding explanations as communicative acts, in particular the effects of the reliability of the source on confidence. I have suggested direct implications of the findings for both the psychological research and the research in AI as well as a number of avenues for further research. As one of the contentions of this thesis is that explanations and arguments should be studied in tandem rather than in isolation, any future research on explanation should consider arguments and their relations to explanations.

Appendices

A

Calculations and experimental material used in studies in Chapter 2

A.1 Explaining away with one or more inhibitory causes

Here I show that Inequality 2.2 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:

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

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, I write again Inequality 2.2:

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

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 2.2 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 2.2 remains satisfied.

A.2 Normative predictions based on data from Rottman and Hastie (2016)

Here I 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 I 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 $P(\mathrm{C_i} \mid {\sim}\mathrm{C_j}) = .35$.

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

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

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.

A.3 The decomposition conditions for an explaining away CBN with two effects

I adopt the following convention: $\overline{a} = 1 - a$.

Theorem A.1. *$P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) = P_2(\mathrm{C}_1^* \mid \mathrm{E}_2)$ if and only if (i) $\alpha_1\,\delta_1 = \beta_1\,\gamma_1$ or (ii) $\alpha_2 = \beta_2$ and $\gamma_2 = \delta_2$.*

Proof.

$$
\begin{aligned}
P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) &= \frac{P_1(\mathrm{C}_1, \mathrm{E}_1, \mathrm{E}_2)}{P_1(\mathrm{E}_1, \mathrm{E}_2)} \\
&= \frac{P_1(\mathrm{C}_1) \sum_{C_2} P_1(\mathrm{E}_1 \mid \mathrm{C}_1, C_2)\, P_1(\mathrm{E}_2 \mid \mathrm{C}_1, C_2)\, P_1(C_2)}{\sum_{C_1, C_2} P_1(\mathrm{E}_1 \mid C_1, C_2)\, P_1(\mathrm{E}_2 \mid C_1, C_2)\, P_1(C_1)\, P_1(C_2)} \\
&= \frac{A_1}{A_1 + A_2} \\
A_1 &:= c_1\,(\alpha_2\,\alpha_1\,c_2 + \beta_2\,\beta_1\,\overline{c_2}) \\
A_2 &:= \overline{c_1}\,(\gamma_2\,\gamma_1\,c_2 + \delta_2\,\delta_1\,\overline{c_2}) \\
P_2(\mathrm{C}_1^* \mid \mathrm{E}_2) &= \frac{P_2(\mathrm{C}_1^*, \mathrm{E}_2)}{P_2(\mathrm{E}_2)} \\
&= \frac{P_2(\mathrm{C}_1^*) \sum_{C_2} P_2(\mathrm{E}_2 \mid \mathrm{C}_1^*, C_2^*)\, P_1(C_2^*)}{\sum_{C_1^*, C_2^*} P_2(\mathrm{E}_2 \mid C_1^*, C_2^*)\, P_2(C_1^*) P_2(C_2^*)} \\
&= \frac{P_1(\mathrm{C}_1 \mid \mathrm{E}_1) \sum_{C_2} P_1(\mathrm{E}_2 \mid \mathrm{C}_1, C_2)\, P_1(C_2 \mid \mathrm{E}_1)}{\sum_{C_1, C_2} P_1(\mathrm{E}_2 \mid C_1, C_2)\, P_1(C_1 \mid \mathrm{E}_1) P_1(C_2 \mid \mathrm{E}_1)} \\
&= \frac{B_1}{B_1 + B_2} \\
B_1 &:= c_1\,(\alpha_1\,c_2 + \beta_1\,\overline{c_2})\cdot
\end{aligned}
$$

$$\cdot\ [\alpha_2\, c_2\, (\alpha_1\, c_1 + \gamma_1\, \overline{c_1}) + \beta_2\, \overline{c_2}\, (\beta_1\, c_1 + \delta_1\, \overline{c_1})]$$

$$B_2 := \overline{c_1}\, (\gamma_1\, c_2 + \delta_1\, \overline{c_2}) \cdot$$

$$\cdot\ [\gamma_2\, c_2\, (\alpha_1\, c_1 + \gamma_1\, \overline{c_1}) + \delta_2\, \overline{c_2}\, (\beta_1\, c_1 + \delta_1\, \overline{c_1})]$$

Let $\Delta_1 := P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) - P_2(\mathrm{C}_1^* \mid \mathrm{E}_2)$. Then

$$\begin{aligned}
\Delta_1 &= \frac{A_1\, (B_1 + B_2) - B_1\, (A_1 + A_2)}{(A_1 + A_2)\, (B_1 + B_2)} \\
&= \frac{A_1\, B_1 + A_1\, B_2 - A_1\, B_1 - A_2\, B_1}{P_1(\mathrm{E}_1, \mathrm{E}_2)\, P_2(\mathrm{E}_2)} = \frac{A_1\, B_2 - A_2\, B_1}{P_1(\mathrm{E}_1, \mathrm{E}_2)\, P_2(\mathrm{E}_2)} \\
&= \frac{c_1\, \overline{c_1}\, c_2\, \overline{c_2}\, (\alpha_1\, \delta_1 - \beta_1\, \gamma_1)\ [G_1 + G_2]}{P_1(\mathrm{E}_1, \mathrm{E}_2)\, P_2(\mathrm{E}_2)}.
\end{aligned}$$

$$G_1 := (\gamma_2 - \delta_2)\, c_1\, (\alpha_2\, \alpha_1\, c_2 + \beta_2\, \beta_1\, \overline{c_2})$$

$$G_2 := (\alpha_2 - \beta_2)\, \overline{c_1}\, (\gamma_2\, \gamma_1\, c_2 + \delta_2\, \delta_1\, \overline{c_2})$$

❑

Using a similar proof strategy one can show that: (a) $P_1(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2) = P_2(\mathrm{C}_2^* \mid \mathrm{E}_2)$ if and only if $\alpha_1\, \delta_1 = \beta_1\, \gamma_1$ or (ii) $\alpha_2 = \gamma_2$ and $\beta_2 = \delta_2$; (b) $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) = P_3(\mathrm{C}_1^* \mid \mathrm{E}_1)$ if and only if (i) $\alpha_2\, \delta_2 = \beta_2\, \gamma_2$ or (ii) $\alpha_1 = \beta_1$ and $\gamma_1 = \delta_1$; and (c) $P_1(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2) = P_3(\mathrm{C}_2^* \mid \mathrm{E}_1)$ if and only if (i) $\alpha_2\, \delta_2 = \beta_2\, \gamma_2$ or (ii) $\alpha_1 = \gamma_1$ and $\beta_1 = \delta_1$ (proofs omitted).

It follows then that $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) = P_2(\mathrm{C}_1^* \mid \mathrm{E}_2) = \mathrm{P}_3(\mathrm{C}_1^* \mid \mathrm{E}_1)$ if (1) $\alpha_1\, \delta_1 = \beta_1\, \gamma_1$ and $\alpha_2\, \delta_2 = \beta_2\, \gamma_2$, or (2) $\alpha_1 = \beta_1$ and $\gamma_1 = \delta_1$, or (3) $\alpha_2 = \beta_2$ and $\gamma_2 = \delta_2$. Similarly, $P_1(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2) = P_2(\mathrm{C}_2^* \mid \mathrm{E}_2) = \mathrm{P}_3(\mathrm{C}_2^* \mid \mathrm{E}_1)$ if (1) $\alpha_1\, \delta_1 = \beta_1\, \gamma_1$ and $\alpha_2\, \delta_2 = \beta_2\, \gamma_2$, or (2) $\alpha_1 = \gamma_1$ and $\beta_1 = \delta_1$, or (3) $\alpha_2 = \gamma_2$ and $\beta_2 = \delta_2$. Therefore, the order is not important and one can decompose a full CBN in smaller ones while preserving the same probability distributions if (1) $\alpha_1\, \delta_1 = \beta_1\, \gamma_1$ and $\alpha_2\, \delta_2 = \beta_2\, \gamma_2$; or (2) $\alpha_1 = \beta_1$, $\gamma_1 = \delta_1$, $\alpha_2 = \gamma_2$, and $\beta_2 = \delta_2$; or (3) $\alpha_2 = \beta_2$, $\gamma_2 = \delta_2$, $\alpha_1 = \gamma_1$, and $\beta_1 = \delta_1$; or (4) $\alpha_1 = \beta_1 = \gamma_1 = \delta_1$; or (5) $\alpha_2 = \beta_2 = \gamma_2 =$

δ_2. (4) and (5) make E_1 and E_2 respectively fully undiagnostic with respect to C_1 and C_2, which violates the faithfulness condition (see Neapolitan, 2003). (1) implies that C_1 and C_2 are conditionally independent given E_1 and that they are also conditionally independent given E_2, that is, learning E_1 makes C_1 and C_2 independent and learning E_1 makes C_1 and C_2 independent. (2) and (3) both entail (1) and are more specific versions of (1).

A.4 Order effects with mutually exclusive and exhaustive causes

Here I show that there are no order effects when the causes are mutually exclusive and exhaustive, i.e. when $P(C_1, C_2) = 0$ and $P(C_1) + P(C_2) = 1$. I model mutually exclusive and exhaustive causes with one node, C, that has two values: C_1 and C_2.

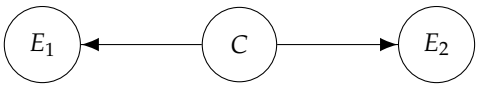


Figure A.1: CBN with mutually exclusive and exhaustive causes

$$\begin{aligned} P_4(C = \mathrm{C}_1) = c \quad &, \quad P_4(C = \mathrm{C}_2) = \overline{c} \\ P_4(\mathrm{E}_1 \mid \mathrm{C}_1) = \alpha_1 \quad &, \quad P_4(\mathrm{E}_1 \mid \mathrm{C}_2) = \beta_1 \\ P_4(\mathrm{E}_2 \mid \mathrm{C}_1) = \alpha_2 \quad &, \quad P_4(\mathrm{E}_2 \mid \mathrm{C}_2) = \beta_2 \end{aligned} \tag{A.1}$$

Splitting the CBN from Figure A.1 we get two CBNs:

$$P_5(C = \mathrm{C}_1) = c \quad , \quad P_5(C = \mathrm{C}_2) = \overline{c}$$

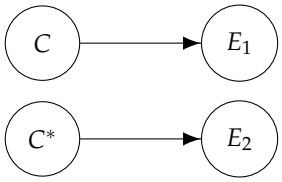


Figure A.2: 'Split' CBN from E_1 to E_2

$$\begin{aligned} P_5(C^* = \mathrm{C}_1^*) = P_4(\mathrm{C}_1 \mid \mathrm{E}_1) \quad , \quad P_5(C^* = \mathrm{C}_2^*) = P_4(\mathrm{C}_2 \mid \mathrm{E}_1) \\ P_5(\mathrm{E}_1 \mid \mathrm{C}_1) = \alpha_1 \quad , \quad P_5(\mathrm{E}_1 \mid \mathrm{C}_2) = \beta_1 \\ P_5(\mathrm{E}_2 \mid \mathrm{C}_1^*) = \alpha_2 \quad , \quad P_5(\mathrm{E}_2 \mid \mathrm{C}_2^*) = \beta_2 \end{aligned} \tag{A.2}$$

Theorem A.2. $P_4(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) = P_5(\mathrm{C}_1^* \mid \mathrm{E}_2)$ *when* $P_{4,5}(C_1^{(*)}, C_2^{(*)}) = 0$ *and* $P_{4,5}(C_1^{(*)}) + P_{4,5}(C_2^{(*)}) = 1$.

Proof.

$$\begin{aligned} P_4(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) &= \frac{P_4(\mathrm{C}_1)\, P_4(\mathrm{E}_1 \mid \mathrm{C}_1)\, P_4(\mathrm{E}_2 \mid \mathrm{C}_1)}{\sum_C P_4(C)\, P_4(\mathrm{E}_1 \mid C)\, P_4(\mathrm{E}_2 \mid C)} \\ &= \frac{c\,\alpha_1\,\alpha_2}{c\,\alpha_1\,\alpha_2 + \bar{c}\,\beta_1\,\beta_2} \\ P_5(\mathrm{C}_1^* \mid \mathrm{E}_2) &= \frac{P_5(\mathrm{C}_1^*)\, P_5(\mathrm{E}_2 \mid \mathrm{C}_1^*)}{\sum_{C^*} P_5(C^*)\, P_5(\mathrm{E}_2 \mid C^*)} \\ &= \frac{P_4(\mathrm{C} \mid \mathrm{E}_1)\, P_4(\mathrm{E}_2 \mid \mathrm{C})}{\sum_C P_4(C \mid \mathrm{E}_1)\, P_4(\mathrm{E}_2 \mid C)} \\ &= \frac{J\,\alpha_2}{J\,\alpha_2 + (1-J)\,\beta_2} \\ J &:= \frac{c\,\alpha_1}{c\,\alpha_1 + \bar{c}\,\beta_1} \end{aligned}$$

Let $\Delta_2 := P_4(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) - P_5(\mathrm{C}_1^* \mid \mathrm{E}_2)$. Then

$$\Delta_2 = \frac{c\,\alpha_1\,\alpha_2\,\beta_2\,\left[1 - \frac{c\,\alpha_1 + \bar{c}\,\beta_1}{c\,\alpha_1 + \bar{c}\,\beta_1}\right]}{(c\,\alpha_1\,\alpha_2 + \bar{c}\,\beta_1\,\beta_2)(J\,\alpha_2 + (1-J)\,\beta_2)} = 0$$

❑

Since $P_4(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2) = 1 - P_4(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2)$ and $P_5(\mathrm{C}_2^* \mid \mathrm{E}_2) = 1 - \mathrm{P}_5(\mathrm{C}_1^* \mid \mathrm{E}_2)$, then given Theorem A.2 it also true that $P_4(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2) = P_5(\mathrm{C}_2^* \mid \mathrm{E}_2)$.

Similarly we get that $P_4(C_1 \mid E_1, E_2) - P_6(C_1^* \mid E_1) = 0$ and $P_4(C_2 \mid E_1, E_2) - P_6(C_2^* \mid E_1) = 0$ (proofs omitted).

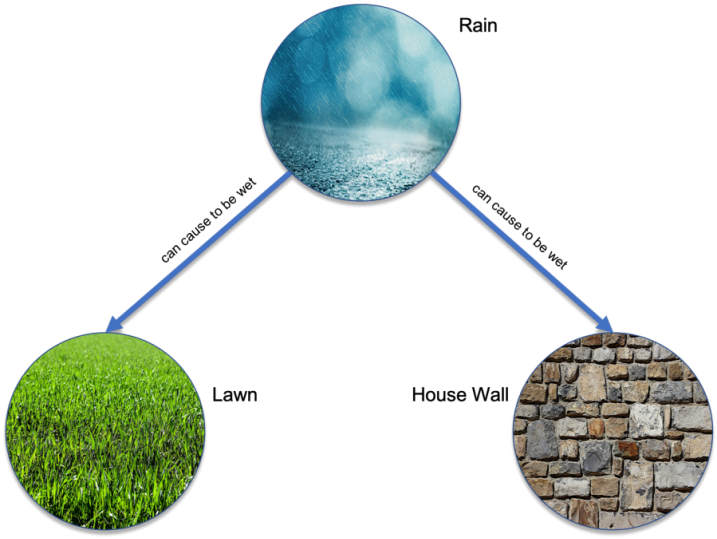
A.5 Stimuli used in Experiment 4

- *The full algebra condition. Step-by-step order 1 (Step-by-step order 2) {All-at-once}*. The text in the square brackets [] did not appear in the materials presented to the participants. Rather it was added here to aid the explanation of the role of each part of the experimental design.

  One morning you plan to do some gardening that day in a garden just a mile away from where you live. However, you won't be able to proceed with your plan and you will have to postpone gardening for another day if it had rained last night. You slept tightly last night so do not remember hearing any rain, but you remember the weather forecast saying that there was a **15% chance** of rain overnight.

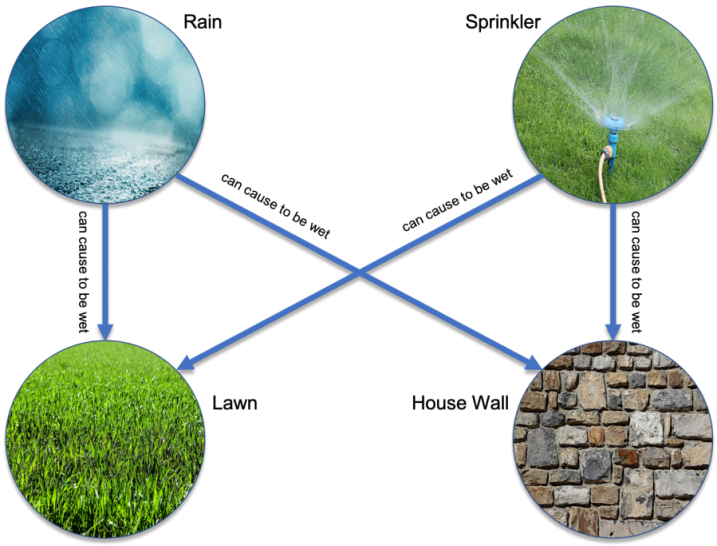
  You head outside to check for signs of rain. You decide to check your lawn and your exterior house wall which sometimes gets wet due to rain.

  The situation is illustrated below:



On the way outside you realize you have to be careful in judging whether it rained last night on the basis of whether the lawn is wet and/or whether the house wall is wet, since you have a lawn sprinkler that sometimes accidentally turns on overnight. The sprinkler can also wet the house wall and turns off early in the morning (before you wake up) if left on overnight due to the water supply to the sprinkler being automatically cut. There is a **15% chance** that your sprinkler accidentally turned on overnight.

The complete situation is illustrated below:



[Questions about priors:]

Q1. How confident are you that it **rained** overnight?

Q2. How confident are you that **the sprinkler turned on** overnight?

From your experience you know that if it rained overnight, but the sprinkler *did not* turn on overnight, then there is a high **70% chance** that the lawn is wet, and a low **20% chance** that the house wall is wet in the morning.

The same values hold if the sprinkler turned on, but it *did not* rain overnight: in this case there is also a high **70% chance** that the lawn is wet, and a low **20% chance** that the house wall is wet in the morning.

If it *both* rained overnight and the sprinkler turned on overnight, then there is an extremely high **99% chance** that the lawn is wet and a moderate **60% chance** that the house wall is wet in the morning.

In case it *neither* rained *nor* the sprinkler turned on overnight, then both the lawn and the house wall are **definitely dry** in the morning.

[Comprehension questions:]

What is the chance that the lawn is wet if it **did not rain** but **the sprinkler turned on** overnight?

What is the chance that the house wall is wet if it **did not rain** but **the sprinkler turned on** overnight?

What is the chance that the lawn is wet if **the sprinkler did not turn** but it **rained** overnight?

What is the chance that the house wall is wet if **the sprinkler did not turn on** but it **rained** overnight?

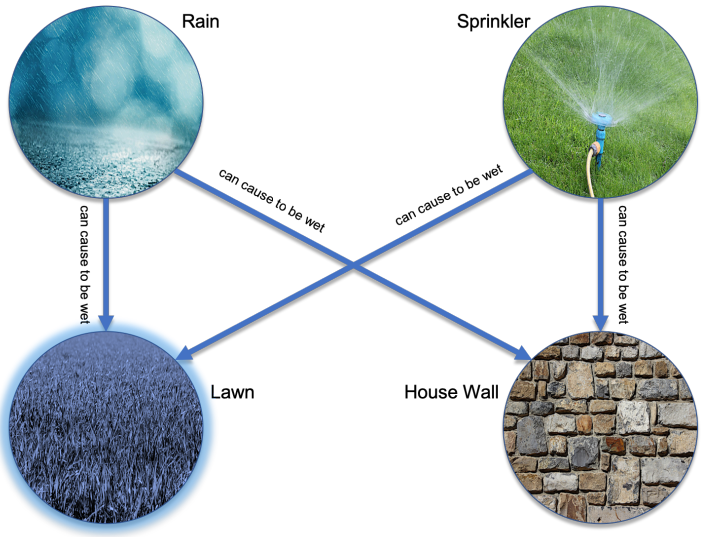
What is the chance that the lawn is wet if it **did not rain** and **the sprinkler did not turn on** overnight?

What is the chance that the house wall is wet if it **did not rain** and **the sprinkler did not turn on** overnight?

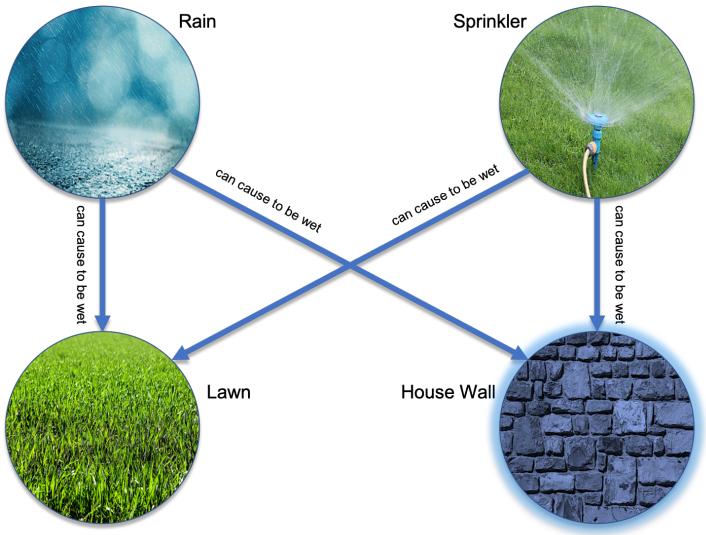
What is the chance that the lawn is wet if it **rained** and **the sprinkler turned on** overnight?

What is the chance that the house wall is wet if it **rained** and **the sprinkler turned on** overnight?

YOU STEP OUTSIDE AND YOU FIRST CHECK THE LAWN. YOU FIND OUT THAT THE LAWN IS WET. You still do not know whether the house wall is wet.



(YOU STEP OUTSIDE AND YOU FIRST LOOK AT THE HOUSE WALL. YOU FIND OUT THAT THE HOUSE WALL IS WET. You still do not know whether the lawn wall is wet.)



{YOU HEAD OUTSIDE AND YOU CHECK BOTH THE LAWN AND THE HOUSE WALL. YOU FIND OUT THAT THE LAWN IS WET AND THAT THE HOUSE WALL IS WET. [Test questions Q3 and Q4

were not asked in this all-at-once condition. Only test questions Q5 and Q6 were asked in this condition.]}

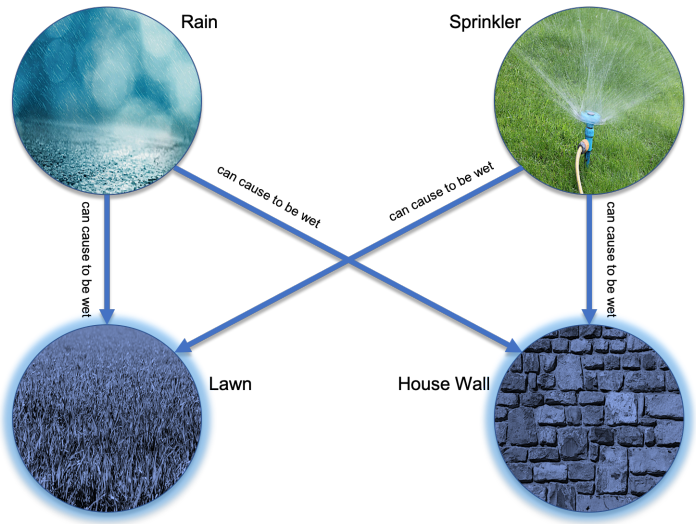
[Test questions 1:]

Q3. How confident are you that it **rained** overnight now that you know that the lawn is wet (the house wall is wet)?

Q4. How confident are you that **the sprinkler turned on** overnight now that you know that the lawn is wet (the house wall is wet)?

YOU THEN LOOK AT THE HOUSE WALL AND FIND OUT THAT THE HOUSE WALL IS ALSO WET.

(YOU THEN CHECK THE LAWN AND FIND OUT THAT THE LAWN IS ALSO WET.)



[Test questions 2:]

Q5. How confident are you that it **rained** overnight now that you know that *both* the lawn *and* the house wall are wet?

Q6. How confident are you that **the sprinkler turned on** overnight now that you know that *both* the lawn *and* the house wall are wet?

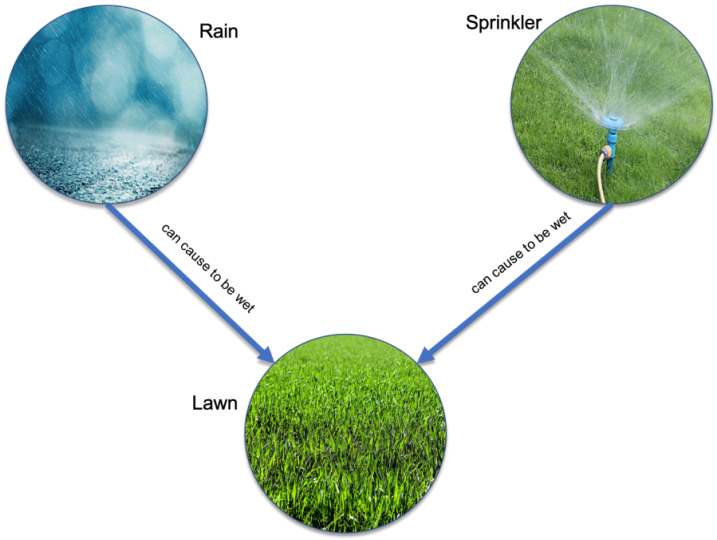
- *The sequential algebra condition. Step-by-step order 1 (Step-by-step order 2) {All-at-once}*

One morning you plan to do some gardening that day in a garden just a mile away from where you live. However, you won't be able to proceed with your plan and you will have to postpone gardening for another day if it had rained last night. You slept tightly last night so do not remember hearing any rain, but you remember the weather forecast saying that there was a **15% chance** of rain overnight.

You head outside to check for signs of rain. You decide to check your lawn (the house wall which sometimes gets wet due to rain).

On the way outside you realize you have to be careful in judging whether it rained last night on the basis of whether the lawn (the house wall) is wet, since you have a lawn sprinkler that sometimes accidentally turns on overnight. The sprinkler (can also wet the house wall and) turns off early in the morning before you wake up if left on overnight due to the water supply to the sprinkler being automatically cut. There is a **15% chance** that your sprinkler accidentally turned on overnight.

The situation is illustrated below:



([In Step-by-step order 2 participants were presented with the illustration like the one right after Test question 1 below.])

[Questions about priors. The same questions as in the full algebra condition.]

From your experience you know that if it rained overnight, but the sprinkler *did not* turn on overnight, then there is a high **70% chance** that the lawn is wet.

The same values hold if the sprinkler turned on, but it *did not* rain overnight: in this case there is also a high **70% chance** that the lawn is wet.

If it *both* rained overnight *and* the sprinkler turned on overnight, then there is an extremely high **99% chance** that the lawn is wet.

In case it *neither* rained *nor* the sprinkler turned on overnight, then the lawn is **definitely dry** in the morning.

([In Step-by-step order 2 participants were presented with probabilities related to the house wall.])

[Comprehension questions 1:]

What is the chance that the lawn is wet if it **did not rain** but **the sprinkler turned on** overnight?

What is the chance that the lawn is wet if **the sprinkler did not turn** but it **rained** overnight?

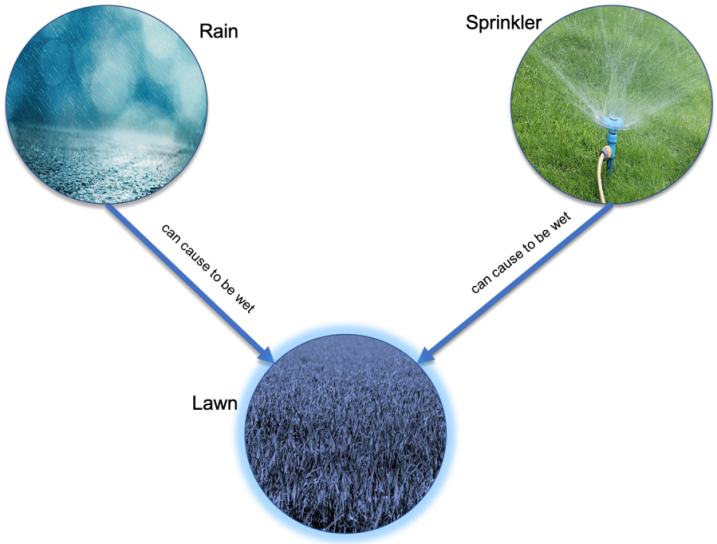
What is the chance that the lawn is wet if it **did not rain** and **the sprinkler did not turn on** overnight?

What is the chance that the lawn is wet if it **rained** and **the sprinkler turned on** overnight?

([In Step-by-step order 2 participants were presented with Comprehension questions 2.])

YOU STEP OUTSIDE AND YOU CHECK THE LAWN. YOU FIND OUT THAT THE LAWN IS WET.

(**YOU STEP OUTSIDE AND LOOK AT THE HOUSE WALL. YOU FIND OUT THAT THE HOUSE WALL IS WET.**)



([In Step-by-step order 2 participants were presented with a picture like the one just before Test questions 2.])

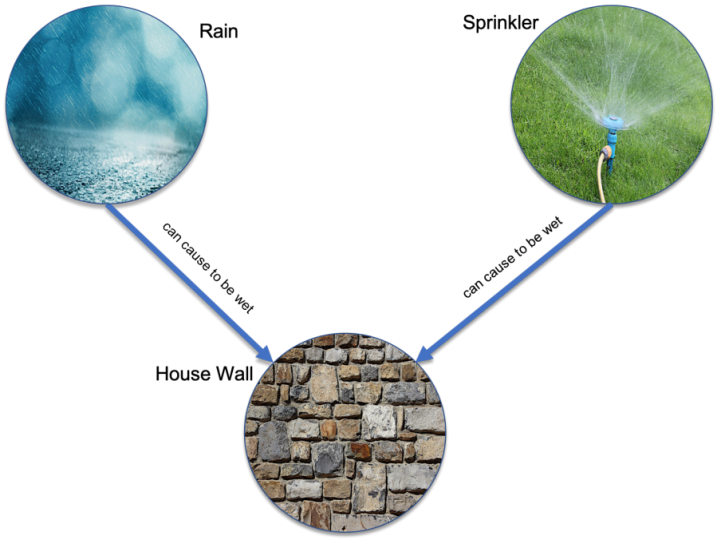
[Test questions 1:]

Q3. How confident are you that it **rained** overnight now that you know that the lawn is wet (the house wall is wet)?

Q4. How confident are you that **the sprinkler turned on** overnight now that you know that the lawn is wet (the house wall is wet)?

To get more information on rain overnight, you decide to also check your exterior house wall (your lawn) that can also sometimes get wet both due to rain and due to the sprinkler being on.

The situations is illustrated bellow:



([In Step-by-step order 2 participants were presented with like the one just before priors.])

From your experience you know that if it rained overnight, but the sprinkler *did not* turn on overnight, then there is a low **20% chance** that the house wall is wet in the morning.

The same values hold if the sprinkler turned on, but it *did not* rain overnight: in this case there is also a low **20% chance** that the house wall is wet in the morning.

If it *both* rained overnight *and* the sprinkler turned on overnight, then there is a moderate **60% chance** that the house wall is wet in the morning.

In case it *neither* rained *nor* the sprinkler turned on overnight, then the house wall is **definitely dry** in the morning.

([In Step-by-step order 2 participant were presented with probabilities related to the lawn.])

[Comprehension questions 2:]

What is the chance that the house wall is wet if it **did not rain** but **the sprinkler turned on** overnight?

What is the chance that the house wall is wet if **the sprinkler did not turn on** but it **rained** overnight?

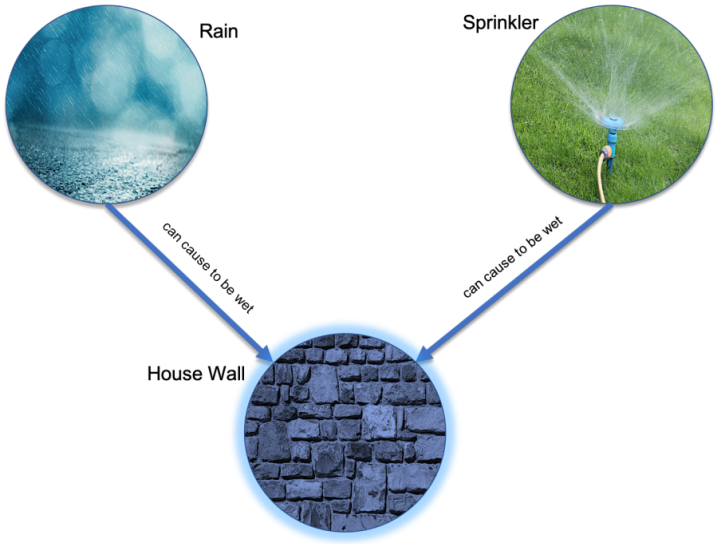
What is the chance that the house wall is wet if it **did not rain** and **the sprinkler did not turn on** overnight?

What is the chance that the house wall is wet if it **rained** and **the sprinkler turned on** overnight?

([In Step-by-step order 2 participant were presented with Comprehension questions 1.])

YOU LOOK AT THE HOUSE WALL AND YOU FIND OUT THAT THE HOUSE WALL IS ALSO WET.

(YOU CHECK THE LAWN AND FIND OUT THAT THE LAWN IS ALSO WET.)



[Test questions 2:]

Q5. How confident are you that it **rained** overnight now that you know that the house wall (the lawn) is also wet?

Q6. How confident are you that **the sprinkler turned on** overnight now that you know that the house wall (the lawn) is also wet?

{[In All-at-one condition people learnt the effects in a sequential manner but the were told about potential existence of both effect they were told that both effects had occurred and they were asked only Test questions 2.]}

A.6 Ratio of the posterior odds for the full and the split model

In this section I show how one can represent the predictions of the full and the split model as a ratio of odds that simplifies to likelihoods. For the full model

(over which a probability distribution P_7 is defined) we have the following ratio of the odds ($i, j \in \{1, 2\}$):

$$\begin{aligned}
\frac{\dfrac{P_7(\mathrm{C_i} \mid \mathrm{E_i})}{P_7(\mathrm{C_i} \mid \mathrm{E_i}, \mathrm{E_j})}}{\dfrac{P_7(\mathrm{C_j} \mid \mathrm{E_i})}{P_7(\mathrm{C_j} \mid \mathrm{E_i}, \mathrm{E_j})}} &= \frac{P_7(\mathrm{C_i} \mid \mathrm{E_i}) P_7(\mathrm{C_j} \mid \mathrm{E_i}, \mathrm{E_j})}{P_7(\mathrm{C_j} \mid \mathrm{E_i}) P_7(\mathrm{C_i} \mid \mathrm{E_i}, \mathrm{E_j})} = \frac{\dfrac{P_7(\mathrm{E_i} \mid \mathrm{C_i}) P_7(\mathrm{C_i})}{P_7(\mathrm{E_i})} \dfrac{P_7(\mathrm{E_i}, \mathrm{E_j} \mid \mathrm{C_j}) P_7(\mathrm{C_j})}{P_7(\mathrm{E_i}, \mathrm{E_j})}}{\dfrac{P_7(\mathrm{E_i} \mid \mathrm{C_j}) P_7(\mathrm{C_j})}{P_7(\mathrm{E_i})} \dfrac{P_7(\mathrm{E_i}, \mathrm{E_j} \mid \mathrm{C_i}) P_7(\mathrm{C_i})}{P_7(\mathrm{E_i}, \mathrm{E_j})}} \\
&= \frac{P_7(\mathrm{E_i} \mid \mathrm{C_i})}{P_7(\mathrm{E_i} \mid \mathrm{C_j})} \frac{P_7(\mathrm{E_i}, \mathrm{E_j} \mid \mathrm{C_j})}{P_7(\mathrm{E_i}, \mathrm{E_j} \mid \mathrm{C_i})} = \frac{P_7(\mathrm{E_i} \mid \mathrm{C_i})}{P_7(\mathrm{E_i} \mid \mathrm{C_j})} \frac{P_7(\mathrm{E_j} \mid \mathrm{C_j}, \mathrm{E_i}) P_7(\mathrm{E_i} \mid \mathrm{C_j})}{P_7(\mathrm{E_j} \mid \mathrm{C_i}, \mathrm{E_i}) P_7(\mathrm{E_i} \mid \mathrm{C_i})} \\
&= \frac{P_7(\mathrm{E_j} \mid \mathrm{C_j}, \mathrm{E_i})}{P_7(\mathrm{E_j} \mid \mathrm{C_i}, \mathrm{E_i})}
\end{aligned}$$

The probability distribution for the split model depends on the order we learn evidence. There are thus two probability distributions (P_8 and P_9), one for each order. Further, by the definition of the split model we have that $P_{7,8}(\mathrm{C^*_{i,j}}) = P_{7,8}(\mathrm{C_{i,j}} \mid \mathrm{E_{i,j}})$ and $P_{8,9}(\mathrm{E_{i,j}} \mid \mathrm{C^*_{i,j}}) = P_{8,9}(\mathrm{E_{i,j}} \mid \mathrm{C_{i,j}})$.

$$\begin{aligned}
\frac{\dfrac{P_{8,9}(\mathrm{C_i} \mid \mathrm{E_i})}{P_{8,9}(\mathrm{C^*_i} \mid \mathrm{E_j})}}{\dfrac{P_{8,9}(\mathrm{C_j} \mid \mathrm{E_i})}{P_{8,9}(\mathrm{C^*_j} \mid \mathrm{E_j})}} &= \frac{P_{8,9}(\mathrm{C_i} \mid \mathrm{E_i}) P_{8,9}(\mathrm{C^*_j} \mid \mathrm{E_j})}{P_{8,9}(\mathrm{C_j} \mid \mathrm{E_i}) P_{8,9}(\mathrm{C^*_i} \mid \mathrm{E_j})} = \frac{P_{8,9}(\mathrm{C^*_i}) P_{8,9}(\mathrm{C^*_j} \mid \mathrm{E_j})}{P_{8,9}(\mathrm{C^*_j}) P_{8,9}(\mathrm{C^*_i} \mid \mathrm{E_j})} \\
&= \frac{\dfrac{P_{8,9}(\mathrm{C^*_i}) P_{8,9}(\mathrm{E_j} \mid \mathrm{C^*_j}) P_{8,9}(\mathrm{C^*_j})}{P_{8,9}(\mathrm{E_j})}}{\dfrac{P_{8,9}(\mathrm{C^*_j}) P_{8,9}(\mathrm{E_j} \mid \mathrm{C^*_i}) P_{8,9}(\mathrm{C^*_i})}{P_{8,9}(\mathrm{E_j})}} = \frac{P_{8,9}(\mathrm{E_j} \mid \mathrm{C^*_j})}{P_{8,9}(\mathrm{E_j} \mid \mathrm{C^*_i})} \\
&= \frac{P_{8,9}(\mathrm{E_j} \mid \mathrm{C_j})}{P_{8,9}(\mathrm{E_j} \mid \mathrm{C_i})}
\end{aligned}$$

We can see that that the ratio of odds in the full models is equal to the ratio of odds in the split model when E_i and E_j are independent (i.e. when $P(E_j \mid E_j) = P(E_j)$), which is exactly what the split model is assuming. In other words, the split model ratio of odds is a special version of the full model ratio of odds when the two effects are independent.

A.7 Stimuli used in Experiment 5

Two cover stories were used in Experiment 5. The first cover story was the rain and sprinkler cover story used in Experiment 4 (the materials can be found in Appendix A.5) and adapted for Experiment 5. The adaptation is reflected in the features of the second cover story used in Experiment 5. I thus provide the materials of only that cover story.

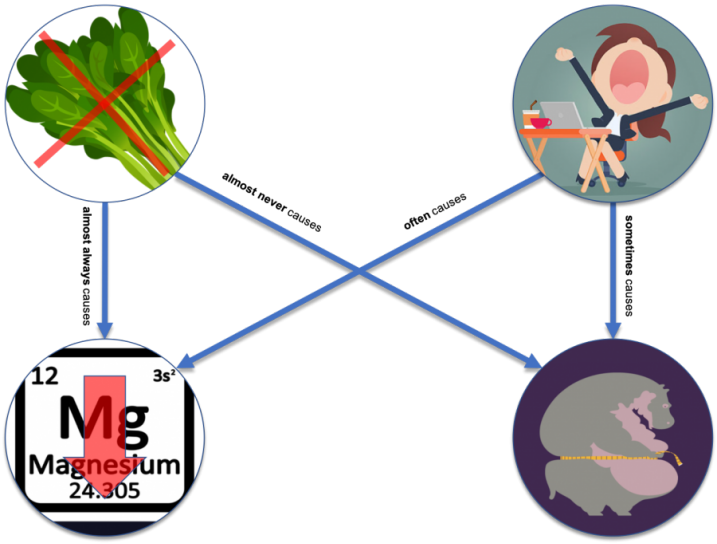
- *The full algebra condition. Step-by-step order 1 (Step-by-step order 2).* The text in the square brackets [] did not appear in the materials presented to the participants. Rather it was added here to aid the explanation of the role of each part of the experimental design.

  Researchers have established that sleep deprivation can be responsible for both magnesium deficiency in the body and obesity. According to the research, sleep deprivation often causes magnesium deficiency in the body and sometimes causes obesity.

  The research has also shown that not regularly eating magnesium-rich

foods (such as leafy green vegetables) almost always causes magnesium deficiency and almost never causes obesity.

The situation is illustrated below:



Your friend Tom, who you haven't seen in a while, is known for really liking vegetables. So, you think it is unlikely that Tom is not regularly consuming magnesium-rich vegetables.

Also, Tom is well-known for having no troubles sleeping and you think it is very unlikely that he is sleep deprived.

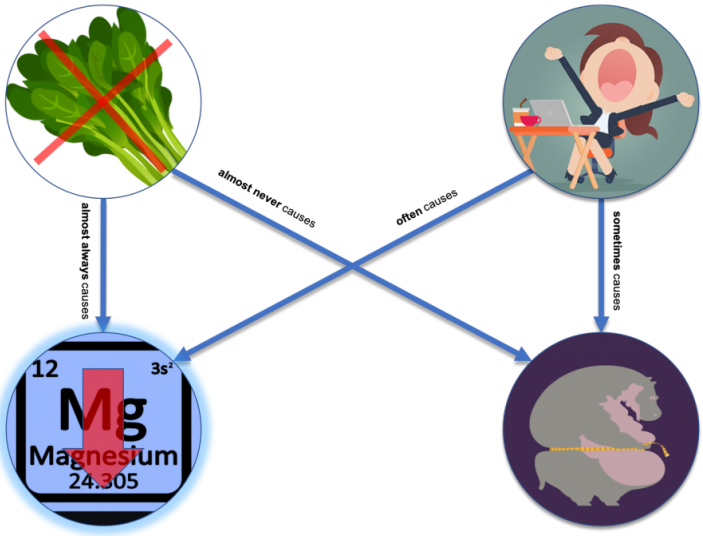
Please answer the following questions using all the information so far.

[Questions about priors:]

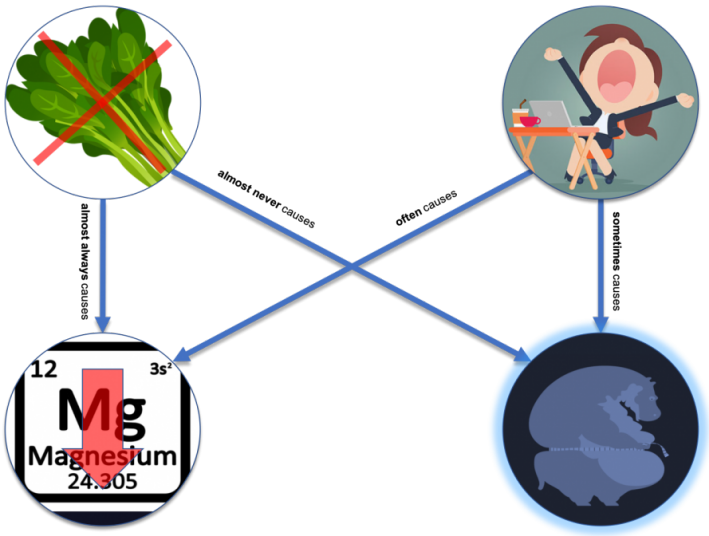
Q1. How confident are you that Tom is not regularly eating magnesium-rich foods?

Q2. How confident are you that Tom is sleep deprived?

YOU CALL TOM AND DURING THE CONVERSATION HE TELLS YOU THAT HE IS MAGNESIUM DEFICIENT. At this point, you still do not know whether he is diagnosed with obesity or not.



(YOU CALL TOM AND DURING THE CONVERSATION HE TELLS YOU THAT HE WAS DIAGNOSED WITH OBESITY. At this point, you still do not know whether he is magnesium deficient or not.)

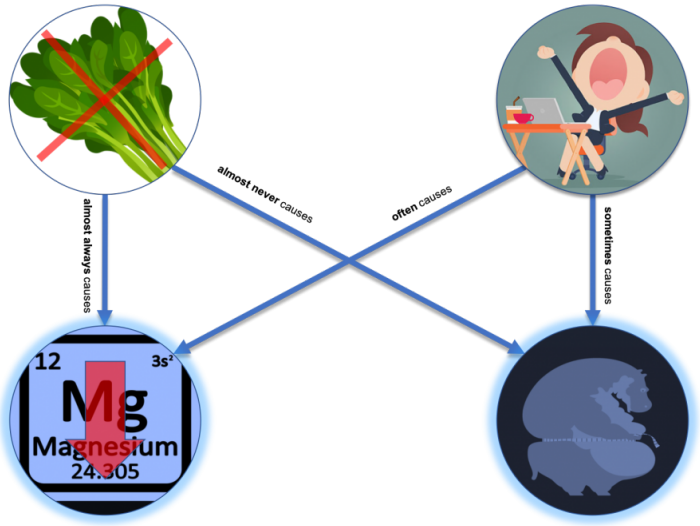


[Test questions 1:]

Q3. How confident are you that Tom is not regularly eating magnesium-rich foods now that you know that he is magnesium deficient (obese)?

Q4. How confident are you that Tom is sleep deprived now that you know that he is magnesium deficient (obese)?

AFTER A WHILE, TOM ALSO TELLS YOU THAT UNFORTUNATELY HE BECAME OBESE (MAGNESIUM DEFICIENT).



[Test questions 2:]

Q5. How confident are you that Tom is not regularly eating magnesium-rich foods now that you know that he is both magnesium deficient and obese?

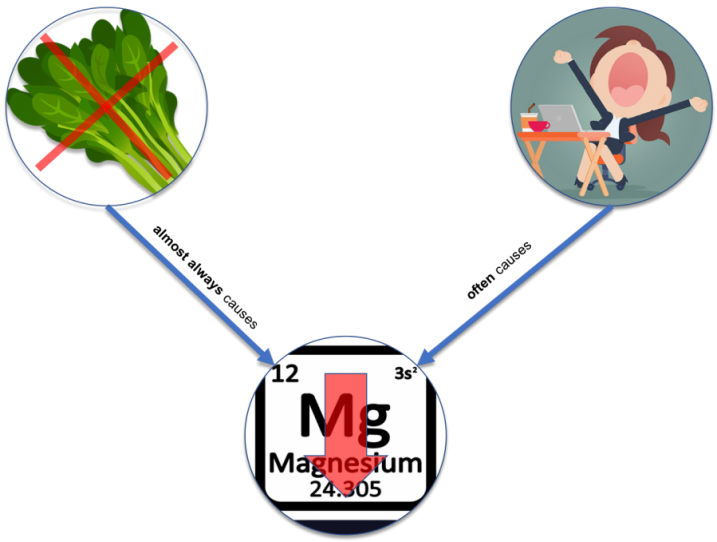
Q6. How confident are you that Tom is sleep deprived now that you know that he is both magnesium deficient and obese?

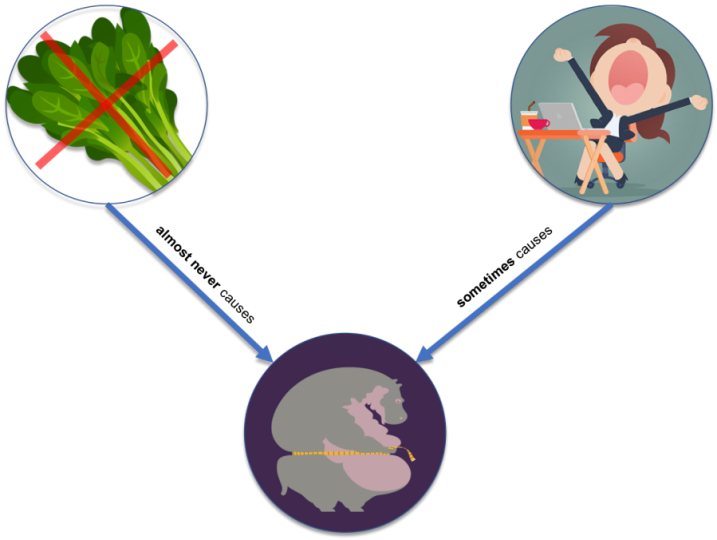
- *The sequential algebra condition. Step-by-step order 1 (Step-by-step order 2)*

Researchers have established that sleep deprivation can be responsible for magnesium deficiency in the body (obesity). According to the research, sleep deprivation often (sometimes) causes magnesium deficiency in the body (obesity).

The research has also shown that not regularly eating magnesium-rich foods (such as leafy green vegetables) almost always (almost never) causes magnesium deficiency (obesity).

The situation is illustrated below:





Your friend Tom, who you haven't seen in a while, is known for really liking vegetables. So, you think it is unlikely that Tom is not regularly consuming magnesium-rich vegetables.

Also, Tom is well-known for having no troubles sleeping and you think it is very unlikely that he is sleep deprived.

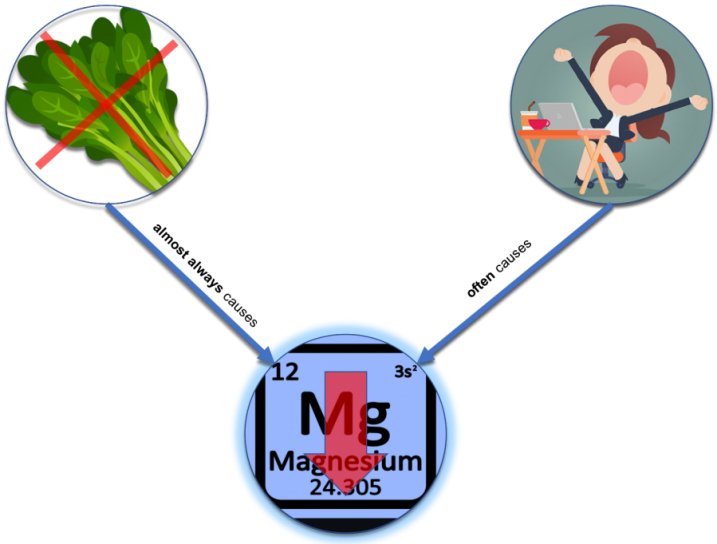
Please answer the following questions using all the information so far.

[Questions about priors:]

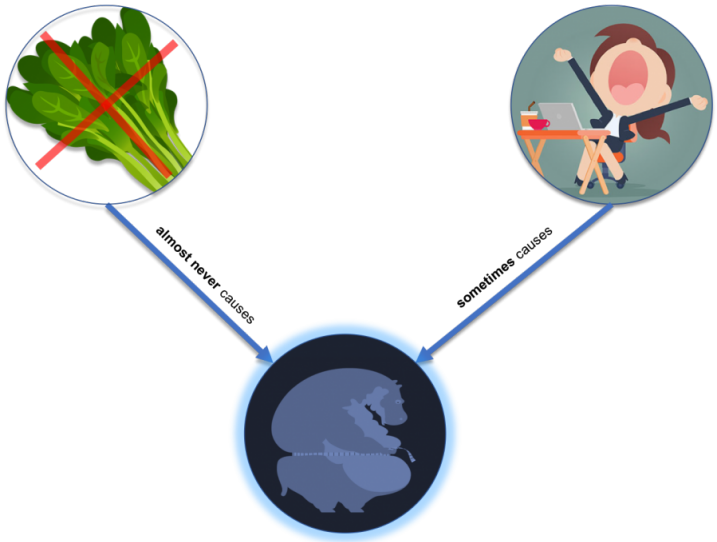
Q1. How confident are you that Tom is not regularly eating magnesium-rich foods?

Q2. How confident are you that Tom is sleep deprived?

YOU CALL TOM AND DURING THE CONVERSATION HE TELLS YOU THAT HE IS MAGNESIUM DEFICIENT.



(YOU CALL TOM AND DURING THE CONVERSATION HE TELLS YOU THAT HE WAS DIAGNOSED WITH OBESITY.)



[Test questions 1:]

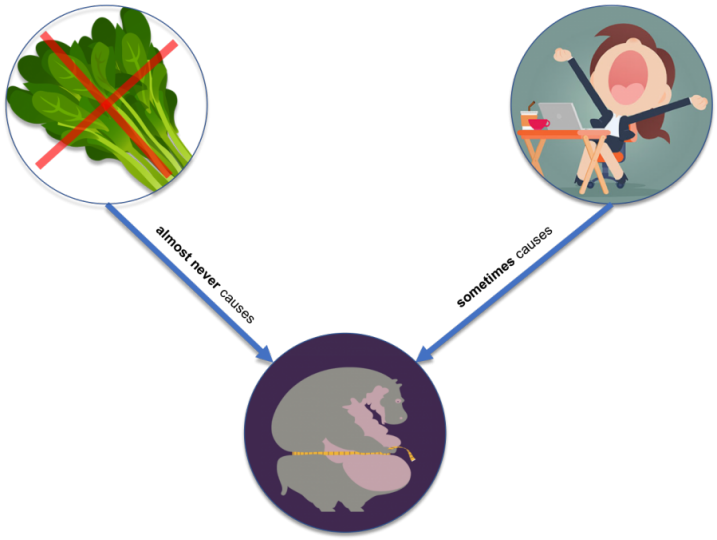
Q3. How confident are you that Tom is not regularly eating magnesium-rich foods now that you know that he is magnesium deficient (obese)?

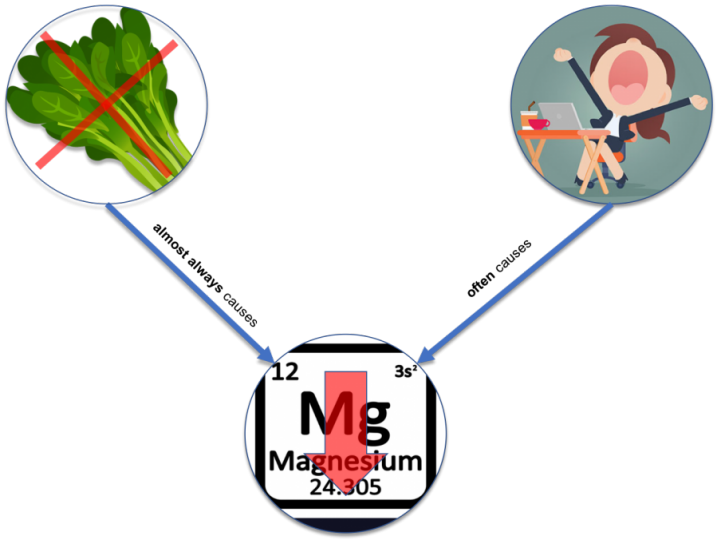
Q4. How confident are you that Tom is sleep deprived now that you know that he is magnesium deficient (obese)?

In the meantime you remembered research on obesity (magnesium deficiency). According to that research, sleep deprivation sometimes (often) causes obesity (magnesium deficiency).

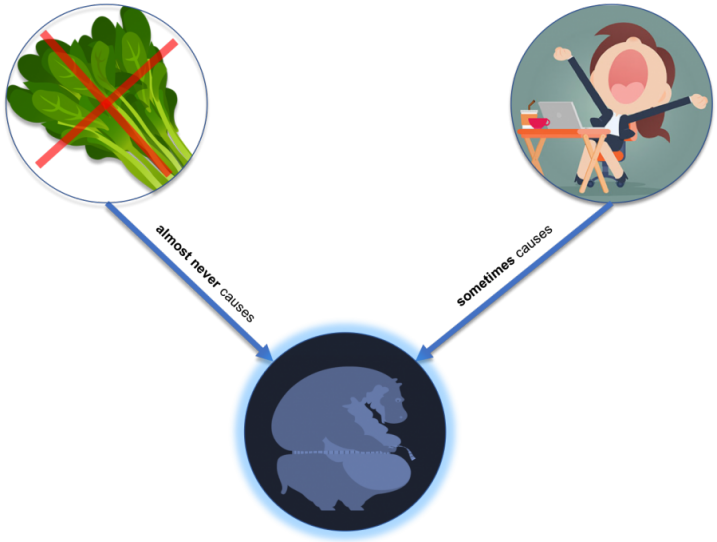
The research has also shown that not regularly eating magnesium-rich foods almost never (almost always) causes obesity (magnesium deficiency).

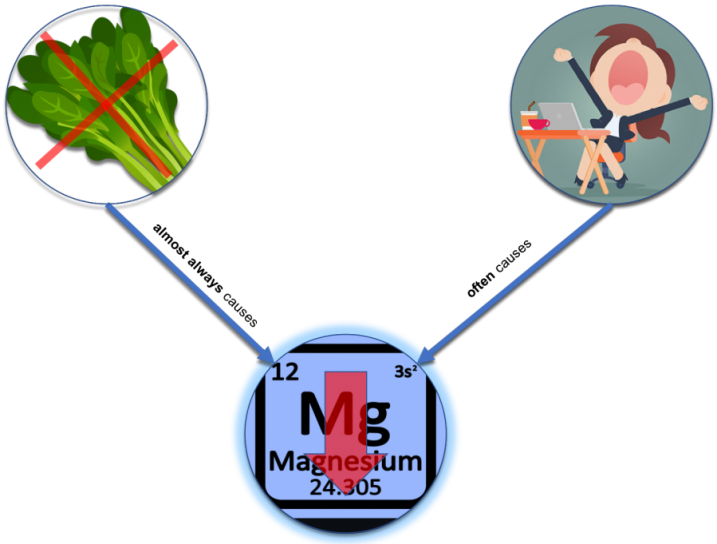
The situations is illustrated bellow:





IN ANOTHER CONVERSATION TOM ALSO TELLS YOU THAT UNFORTUNATELY HE BECAME OBESE (MAGNESIUM DEFICIENT).





[Test questions 2:]

Q5. How confident are you that Tom is not regularly eating magnesium-rich foods now that you additionally know that he became obese (is magnesium deficient)?

Q6. How confident are you that Tom is sleep deprived now that you additionally know that he became obese (is magnesium deficient)?

B

Experimental materials used in the case study in Chapter 3

Participants in the case study were told the following:

The aim of this preliminary study is to get a sense of BN experts' intuitions on how they explain to themselves or to other people with some familiarity

with BNs the change in probabilities of certain variable states given some evidence compared to when that evidence was unknown.

Please consider the following four Bayesian networks: wet grass, chest clinic, false barrier, and car diagnosis. Curly brackets, i.e. {}, indicate all evidence that you have and you are supposed to use only the evidence in {} to answer the subsequent query. You can create a copy of this file or create a new one to answer the queries, but in either case please put your name/initials to the file. You will find that some queries are optional, but it's preferred that you answer as many of them as you can, time permitting of course. Please upload your file with answers to the Dropbox folder. It would be great if you could upload the file within next two weeks.

Accompanying these questions are four Netica and AgenaRisk files with the four BNs already parameterized. You should these four files use to answer the queries.

Wet Grass

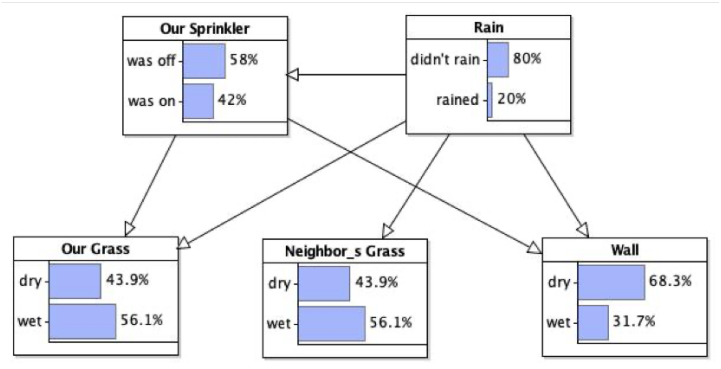


Figure B.1: Wet Grass BN used in the case study.

Evidence: {Neighbour's Grass = wet}

Query:

1. How does the probability of 'Our Sprinkler = was on' change compared to when there was no evidence and why?
2. How does the probability of 'Wall = wet' change compared to when there was no evidence and why?

Evidence: {Wall = wet}
Query:

1. How does the probability of 'Neighbour's Grass = wet' change compared to when there was no evidence and why?
2. How does the probability of 'Rain = rained' change compared to when there was no evidence and why?

Evidence: {Our Grass = wet, Wall = wet}
Query:

1. How does the probability of 'Rain = rained' change compared to when the only available evidence was {Our grass = wet} and why?
2. How does the probability of 'Our Sprinkler = was on' change compared to when the only available evidence was {Our grass = wet} and why?

Evidence: {Our Grass = wet, Wall = wet}
Query:

1. How does the probability of 'Rain = rained' change compared to when there was no evidence and why?

2. How does the probability of 'Our Sprinkler = was on' change compared to when there was no evidence and why?

Optional:

Evidence: {Wall = dry}

Query:

1. How does the probability of 'Neighbour's Grass = wet' change compared to when there was no evidence and why?

2. How does the probability of 'Our Grass = wet' change compared to when there was no evidence and why?

Chest Clinic

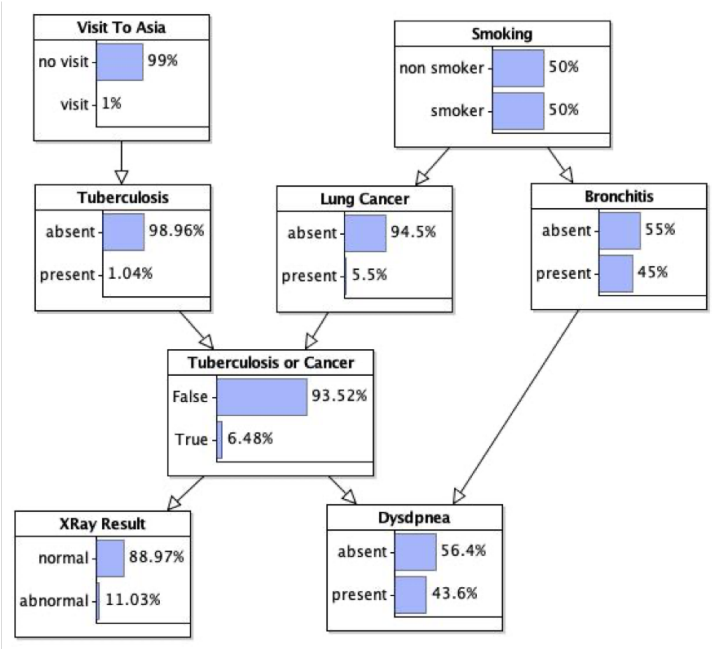


Figure B.2: Chest Clinic BN used in the case study.

Evidence: {XRay Result = abnormal}

Query:

1. How does the probability of 'Lung Cancer = present' change compared to when there was no evidence and why?
2. How does the probability of 'Visit to Asia = visit' change compared to when there was no evidence and why?
3. How does the probability of 'Dyspnea = present' change compared to when there was no evidence and why?

Evidence: {XRay Result = abnormal, Visit to Asia = visit}

Query:

1. How does the probability of 'Lung Cancer = present' change compared to when the only available evidence was XRay Result = abnormal and why?
2. How does the probability of 'Dyspnea = present' change compared to when the only available evidence was XRay Result = abnormal and why?

Evidence: {Smoking = smoker, Bronchitis = present}

Query:

1. How does the probability of 'Lung Cancer = present' change compared to when the only available evidence was Smoking = smoker and why?

   Optional:

2. How does the probability of 'Dyspnea = present' change compared to when the only available evidence was Smoking = smoker and why?

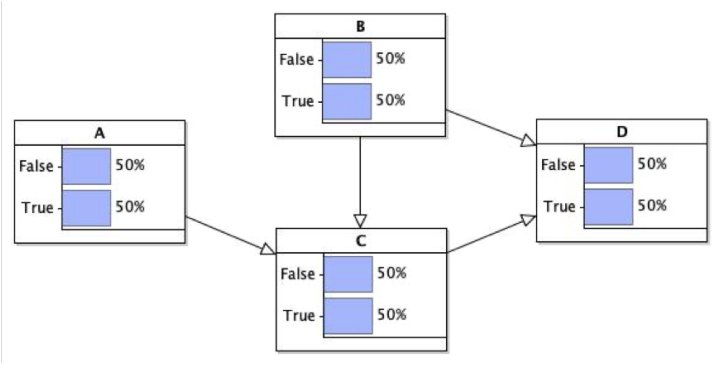


Figure B.3: False Barrier BN used in the case study.

False Barrier

Evidence: {A = true}

Query:

1. How does the probability of 'C = true' change compared to when there was no evidence and why?

2. How does the probability of 'D = true' change compared to when there was no evidence and why?

   Optional:

3. How does the probability of 'B = true' change compared to when there was no evidence and why?

Evidence: {D = true}

Query:

1. How does the probability of 'A = true' change compared to when there was no evidence and why?

2. How does the probability of ‘C = true’ change compared to when there was no evidence and why?

   Optional:

3. How does the probability of ‘B = true’ change compared to when there was no evidence and why?

Evidence: {B = true}
Query:

1. How does the probability of ‘C = true’ change compared to when there was no evidence and why?
2. How does the probability of ‘D = true’ change compared to when there was no evidence and why?

Car Diagnosis

Evidence: {Car Starts = false}
Query:

1. How does the probability of ‘Battery Voltage = dead’ change compared to when there was no evidence and why?
2. How does the probability of ‘Starter System = faulty’ change compared to when there was no evidence and why?

   Optional:

3. How does the probability of ‘Headlights = off’ change compared to when there was no evidence and why?

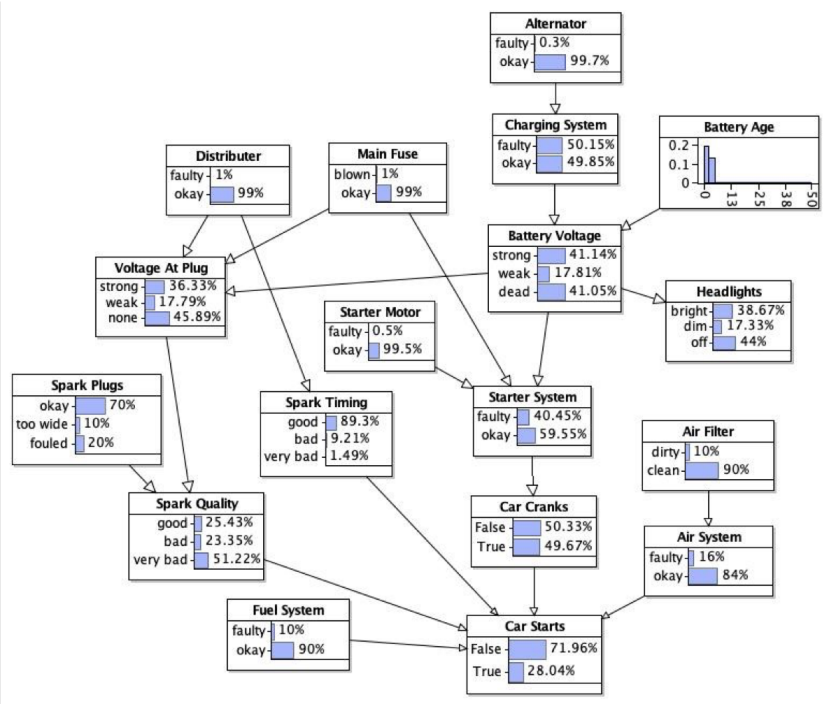


Figure B.4: Car Diagnosis BN used in the case study.

Evidence: {Car Starts = false, Spark Quality = very bad}

Query:

1. How does the probability of 'Battery Voltage = dead' change compared to when the only available evidence was Car Starts = false and why?

2. How does the probability of 'Starter System = faulty' change compared to when the only available evidence was Car Starts = false and why?

3. How does the probability of 'Air System = faulty' change compared to when the only available evidence was Car Starts = false and why?

Evidence: {Car Starts = false, Spark Quality = very bad, Alternator = faulty}

Query:

1. How does the probability of ‘Starter System = faulty’ change compared to when the only available evidence was Car Starts = false, Spark Quality = very bad and why?

2. How does the probability of ‘Starter Motor = faulty’ change compared to when the only available evidence was Car Starts = false, Spark Quality = very bad and why?

3. How does the probability of ‘Air System = faulty’ change compared to when the only available evidence was Car Starts = false, Spark Quality = very bad and why?

   Optional:

4. How does the probability of ‘Spark Plugs = fouled’ change compared to when the only available evidence was Car Starts = false, Spark Quality = very bad and why?

C

Experimental materials used in studies in Chapter 4

C.1 Stimuli used in Experiment 6

The text in parenthesis appeared only on the explanation condition.

1. *The Black Death scenario*:

Millions of people died from the Black Death in the 14th century. **How did the Black Death come to an end?**

One popular belief is that the Black Death subsided mostly through the use of quarantines. (According to this belief, people mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources would traditionally escape to the country, far away from the Black Death-infested cities.)

Q. How confident are you that the Black Death came to an end through the use of quarantines?

2. *Vaccination scenario*:

Vaccination is one of the most common ways to help the immune system. **How do vaccines work?**

It is often thought that vaccines cause the production of antibodies which then strike down viruses. (According to this theory, by administering vaccines one injects weakened versions of viruses which cannot cause an infection. However, the immune cells called 'memory cells' remain in the body. When the body encounters that virus again (now in its harmful version), the memory cells produce antibodies that kill the virus before it's too late.)

Q. How confident are you that vaccines build immunity by causing the production of the antibodies?

3. *China's one-child policy scenario*:

   China had the one-child policy for 35 years. However, China's population has actually grown by about 400 million in these 35 years. **Why has China's population increased if they have had a one-child policy for so long?**

   It is believed that the reason is that the one-child policy did not apply to everyone. (According to this belief, the policy did not apply in rural areas. Also, ethnic minorities were allowed to have more kids. All this resulted in China's population actually growing.)

   Q. How confident are you that China's population has grown because the one-child policy did not apply to everyone?

4. *Ebola scenario*:

   Despite all the modern safety equipment and the fact that Ebola is difficult to transmit, there is still a significant number of medical practitioners who contract Ebola. **Why is that?**

   It is commonly thought that the main reason is the improper removal of the protective gear. (Taking care of someone with Ebola is really difficult. There are body fluids everywhere. So the protective gear is often completely covered with Ebola. Now, when taking off the gear one has to be really careful not to get in contact with the outside of it since they could contract the disease. So even if one has really good protective gear, the improper removal can still lead to contracting Ebola.)

Q. How confident are you that the improper removal of the protective gear is the main reason medical practitioners contract Ebola?

5. *Switzerland in WWII scenario*:

   Switzerland is well-known for its armed neutrality during WWII. **How has Switzerland maintained its armed neutrality during times of conflict like WWII?**

   One popular belief is that Switzerland remained neutral through a combination of military deterrence and economic concessions to Germany. (According to this belief, the Swiss army had a plan to retreat to the mountains in case of an invasion. This would have resulted in Germans having to spend significantly more time and resources in conquering Switzerland. The Swiss army also planned to destroy all major tunnels which would have made any travel from the north to the south of the country practically impossible. On the other hand, the economic cooperation between Switzerland and Germany was high and the Swiss significantly extended credits to Germans. All this contributed to Switzerland successfully remaining neutral.)

   Q. How confident are you that Switzerland maintained its armed neutrality through the combination of military deterrence and economic concession to Germany?

C.2 Stimuli used in Experiments 7a and 7b

The text in parenthesis appeared only on the explanation condition. All participants in Experiment 7a were presented with both Q1 (the confidence question) and Q2 (the reliability question) whereas participants in Experiment 7b were presented either Q1 or Q2, but not both.

1. *The Black Death scenario*:

   Dave and Jimmy are part of a research group investigating devastating pandemics in human history. During a planning meeting they touched upon the Black Death.

   Dave: Millions of people died from the Black Death in the 14th century. I think our research project should in part focus on how the Black Death ended. It may give us some insight into how to deal with future pandemics.

   Jimmy: Yes, I agree. Do you already have an idea regarding how the Black Death came to an end?

   Dave: I think the Black Death subsided mostly through the use of quarantines.

   (**Jimmy**: How so?

   Dave: People mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources would traditionally escape to the country, far away form the Black Death-

infested cities.)

Q1. How confident are you that the Black Death came to an end through the use of quarantines?

Q2. How reliable do you think **Dave** is as a source of information regarding the end of the Black Death?

2. *Vaccination scenario*:

Robert and Michael met to discuss a student project on the ways the immune system can be helped to develop a protection from a disease. During the conversation they touched upon vaccination.

Robert: Vaccination is one of the most common ways to help the immune system. Should we include a section on it in the project?

Michael: Yes, we should. We then need to address how vaccines work. Do you know anything about that?

Robert: I think vaccines build immunity as they contain the non-harmful versions of viruses that cause the production of antibodies which then strike down the harmful versions of viruses.

(**Michael**: How so?

Robert: Well the dead or weakened viruses in the vaccines that are administered cannot cause an infection. However, the immune cells called 'memory cells' remain in the body. When the body encounters that virus again, the memory cells produce antibodies that kill the virus before it's too late.)

Q1. How confident are you that vaccines build immunity by containing the non-harmful versions of viruses that then cause the harmful versions to be eliminated?

Q2. How reliable do you think **Robert** is as a source of information regarding the workings of vaccines?

3. *China's one-child policy scenario*:

Emma and Ben are discussing some of the policies the Chinese government has imposed. During the conversation they touched upon the one-child policy the China had for 35 years.

Emma: The one-child policy seemed like a really extreme way of regulating population size. But China's population has actually grown for about 400 million in the 35 years. Why is that?

Ben: I think the reason is that the one-child policy did not apply to everyone.

(**Emma**: How so?

Ben: Well the policy did not apply in rural areas. Also, ethnic minorities were allowed to have more kids. All this resulted in China's population actually growing.)

Q1. How confident are you that China's population has grown because the one-child policy did not apply to everyone?

Q2. How reliable do you think **Ben** is as a source of information regarding China's one-child policy?

4. *Ebola scenario*:

   Maria is medical practitioner. During one of the conversation with her friend Tom, who is not a doctor, they touched upon reasons doctors contract Ebola.

   Tom: Despite all the modern safety equipment and, as you said, the fact that Ebola is difficult to transmit, there is still a significant number of medical practitioners who contract Ebola. Why is that?

   Maria: I think the main reason is the improper removal of the protective gear.

   (**Tom**: Right. But how does that exactly lead to contracting Ebola?

   Maria: Taking care of someone with Ebola is really difficult. There are body fluids everywhere. So the protective gear is often completely covered with Ebola. Now, when taking off the gear one has to be really careful not to get in contact with the outside of it since they could contract the disease. So even if one has really good protective gear, the improper removal can still lead to contracting Ebola.)

   Q1. How confident are you that the improper removal of the protective gear is the main reason medical practitioners contract Ebola?

   Q2. How reliable do you think **Maria** is as a source of information regarding transmission of Ebola and the proper use of protective equipment?

5. *Switzerland in WWII scenario*:

   Ann and Sarah met to discuss a student project on armed neutrality in World War II. During the conversation they touched upon Switzerland.

Ann: Switzerland is well-known for its neutrality during WWII. Should we include a section on it in the project?

Sarah: Yes, we should. Do you know how Switzerland maintained its armed neutrality during WWII?

Ann: I think they remained neutral through a combination of military deterrence and economic concessions to Germany.

(**Sarah**: Right. Do you know anything more specific?

Ann: The Swiss army had a plan to retreat to the mountains in case of an invasion. This would have resulted in Germans having to spend significantly more time and resources in conquering Switzerland. The Swiss army also planned to destroy all major tunnels which would have made any travel from the north to the south of the country practically impossible. On the other hand, the economic cooperation between Switzerland and Germany was high and the Swiss significantly extended credits to Germans. All this contributed to Switzerland successfully remaining neutral.)

Q1. How confident are you that Switzerland maintained its armed neutrality through the combination of military deterrence and economic concession to Germany?

Q2. How reliable do you think **Ann** is as a source of information regarding Switzerland's armed neutrality during WWII?

C.3 Stimuli used in Experiment 8

The text in parenthesis () appeared only on the explanation condition. The text in square brackets [] appeared only in the low reliability condition and the text in curly brackets {} appeared only in the high reliability condition.

1. *The Black Death scenario*:

   [Dave and Jimmy are high school students who are assigned a student project to find out as much as they can on one of the most devastating pandemics in human history, namely the Black Death.]

   {Dave and Jimmy are senior researchers at a well-established institute for global health and part of the project investigating devastating pandemics in human history. During a planning meeting they touched upon the Black Death.}

   Dave: Millions of people died from the Black Death in the 14th century. I think our project should in part focus on how the Black Death ended.

   Jimmy: Yes, I agree. Do you already have an idea regarding how the Black Death came to an end?

   Dave: I think the Black Death subsided mostly through the use of quarantines.

   (**Jimmy**: How so?

   Dave: People mostly stayed out of the path of infected individuals, rats, and fleas. The uninfected would typically remain in their homes and only leave when it was necessary. Those with the financial resources

would traditionally escape to the country, far away form the Black Death-infested cities.)

Q1. How confident are you that the Black Death came to an end through the use of quarantines?

Q2. How reliable do you think **Dave** is as a source of information regarding the end of the Black Death?

2. *Vaccination scenario*:

[Robert and Michael are subway operators. They both recently became parents and during a coffee break they started talking about vaccination.]

{Michael recently became a father and met with Robert, an immunologist, to discuss the ways the immune system can be helped to develop a protection from a disease. During the conversation they touched upon vaccination.}

Robert: Vaccination is a great way to protect your child from diseases?

Michael: Yes, I am aware of that, but I always wondered how vaccines work.

Robert: Vaccines cause the production of antibodies which then strike down viruses.

(**Michael**: How so?

Robert: By administering vaccines one injects weakened versions of viruses which cannot cause an infection. However, the immune cells called 'memory cells' remain in the body. When the body encounters

that virus again (now in its harmful version), the memory cells produce antibodies that kill the virus before it's too late.)

Q1. How confident are you that vaccines build immunity by causing the production of the antibodies?

Q2. How reliable do you think **Robert** is as a source of information regarding the workings of vaccines?

3. *China's one-child policy scenario*:

[Ben and Emma just started their undergraduate studies in philosophy. They enjoy talking about global issues and during one of their conversations they touched upon the one-child policy China had for 35 years.]

{Ben and Emma are experienced policy-makers who specialize on East Asia. During one of their meetings they discussed policies related to the regulation of population size and they touched upon the one-child policy China had for 35 years.}

Emma: The one-child policy seemed like a really extreme way of regulating population size. But China's population has actually grown for about 400 million in the 35 years. Why is that?

Ben: The reason is that the one-child policy did not apply to everyone.

(**Emma**: How so?

Ben: The policy did not apply in rural areas. Also, ethnic minorities were allowed to have more kids. All this resulted in China's population actually growing.)

Q1. How confident are you that China's population has grown because the one-child policy did not apply to everyone?

Q2. How reliable do you think **Ben** is as a source of information regarding China's one-child policy?

4. *Ebola scenario*:

[Maria and Tom read in the news that more medical practitioners from the Doctors Without Borders team in West Africa contracted Ebola. Although neither Maria nor Tom are medical practitioners, the news attracted their attention and they started discussing it.]

{Maria is a medical practitioner who was part of the Doctors Without Borders team in West Africa treating various epidemic diseases. During one of the conversation with her friend Tom, who is not a doctor, they touched upon reasons doctors contract Ebola.}

Tom: Despite all the modern safety equipment and the fact that Ebola is difficult to transmit, there is still a significant number of medical practitioners who contract Ebola. Why is that?

Maria: The main reason is the improper removal of the protective gear.

(**Tom**: But how does that exactly lead to contracting Ebola?

Maria: Taking care of someone with Ebola is really difficult. There are body fluids everywhere. So the protective gear is often completely covered with Ebola. Now, when taking off the gear one has to be really careful not to get in contact with the outside of it since they could contract the disease. So even if one has really good protective gear, the improper

removal can still lead to contracting Ebola.)

Q1. How confident are you that the improper removal of the protective gear is the main reason medical practitioners contract Ebola?

Q2. How reliable do you think **Maria** is as a source of information regarding transmission of Ebola and the proper use of protective equipment?

5. *Switzerland in WWII scenario*:

[Ann and Sarah are high school students and they have just been assigned a project on armed neutrality in Wolrd War II.]

{Ann and Sarah are history professors who have just been awarded a research grant for a project on armed neutrality in World War II. During their planning meeting they discussed the case of Switzerland.}

Ann: Switzerland is well-known for its armed neutrality during WWII. Should we include a section on it in the project?

Sarah: Yes, we should. Do you know how Switzerland maintained its armed neutrality during WWII?

Ann: They remained neutral through a combination of military deterrence and economic concessions to Germany.

(**Sarah**: Do you know anything more specific?

Ann: The Swiss army had a plan to retreat to the mountains in case of an invasion. This would have resulted in Germans having to spend significantly more time and resources in conquering Switzerland. The Swiss army also planned to destroy all major tunnels which would have made

any travel from the north to the south of the country practically impossible. On the other hand, the economic cooperation between Switzerland and Germany was high and the Swiss significantly extended credits to Germans. All this contributed to Switzerland successfully remaining neutral.)

Q1. How confident are you that Switzerland maintained its armed neutrality through the combination of military deterrence and economic concession to Germany?

Q2. How reliable do you think **Ann** is as a source of information regarding Switzerland's armed neutrality during WWII?

Bibliography

Agrahari, R., Foroushani, A., Docking, T. R., Chang, L., Duns, G., Hudoba, M., … Zare, H. (2018). Applications of Bayesian network models in predicting types of hematological malignancies. *Scientific reports*, *8*(1), 6951.

Ahn, W.-k., Novick, L. R., & Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review*, *10*(3), 746–752.

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403–418.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: the role of explanation in the persistence of discredited information. *Journal of personality and social psychology*, *39*(6), 1037.

Anderson, C. A., & Sechler, E. S. (1986). Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology*, *50*(1), 24.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of personality and social psychology*, *2*(1), 1–9.

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

Antaki, C., & Leudar, I. (1992). Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, *22*(2), 181–194.

Bansal, A., Farhadi, A., & Parikh, D. (2014). Towards transparent systems: Semantic characterization of failure modes. In *European conference on computer vision* (pp. 366–381).

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

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173–1182.

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

Bechlivanidis, C., Lagnado, D., Zemla, J. C., & Sloman, S. (2017). Concreteness and abstraction in everyday explanation. *Psychonomic bulletin & review*, *24*(5), 1451–1464.

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.

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, É. (2012). Non-bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*(7), 1178–1203.

Bhatia, J.-S., & Oaksford, M. (2015). Discounting testimony with the argument ad hominem and a bayesian congruent prior model. *Journal of Experimen-*

tal Psychology: Learning, Memory, and Cognition, *41*(5), 1548.

Bocklisch, F., Bocklisch, S. F., & Krems, J. F. (2012). Sometimes, often, and always: Exploring the vague meanings of frequency expressions. *Behavior Research Methods*, *44*(1), 144–157.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental psychology*, *48*(4), 1156.

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

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive science*, *24*(4), 573–604.

Burton, R. R. (1976). Semantic grammar: An engineering technique for constructing natural language understanding systems.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (Tech. Rep.). Army research lab Aberdeen proving ground MD human research and engineering.

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

Chockalingam, S., Pieters, W., Teixeira, A., & van Gelder, P. (2017). Bayesian network models in cyber security: A systematic review. In *Nordic conference on secure it systems* (pp. 105–122).

Choi, A., Wang, R., & Darwiche, A. (2019). On the relative expressiveness of Bayesian and neural networks. *International Journal of Approximate Reasoning*, *113*, 303–323.

Collins, H., & Evans, R. (2008). *Rethinking expertise*. University of Chicago

Press.

Collins, P., & Hahn, U. (2019). We might be wrong, but we think that hedging doesn't protect your reputation. *Journal of experimental psychology. Learning, memory, and cognition*.

Collins, P., Hahn, U., von Gerber, Y., & Olsson, E. J. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in psychology*, *9*, 18.

Collins, P. J., & Hahn, U. (2018). Communicating and reasoning with verbal probability expressions. *Psychology of Learning and Motivation*, *69*, 67–105.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*(2), 133–152.

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., ... Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455.

Cruz, N., Desai, S. C., Dewitt, S., Hahn, U., Lagnado, D., Liefgreen, A., ... Tešić, M. (2020). Widening access to Bayesian problem solving. *Frontiers in Psychology*, *11*, 660.

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

Dewitt, S., Lagnado, D., & Fenton, N. E. (2018). Updating prior beliefs based on ambiguous evidence. In *Proceedings of the 40th annual conference of the cognitive science society.*

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Douven, I. (2013). Inference to the best explanation, dutch books, and inaccuracy minimisation. *The Philosophical Quarterly*, *63*(252), 428–444.

Douven, I., & Schupbach, J. N. (2015). The role of explanatory considerations in updating. *Cognition*, *142*, 299–311.

Drury, B., Valverde-Rebaza, J., Moura, M.-F., & de Andrade Lopes, A. (2017). A survey of the applications of Bayesian networks in agriculture. *Engineering Applications of Artificial Intelligence*, *65*, 29–42.

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.

Eriksson, K. (2012). The nonsense math effect. *Judgment and decision making*, *7*(6), 746.

Eva, B., Stern, R., & Hartmann, S. (2019). The similarity of causal structure. *Philosophy of Science*, *86*(5), 821–835.

Fallon, C. K., & Blaha, L. M. (2018). Improving automation transparency: Addressing some of machine learning's unique challenges. In *International*

conference on augmented cognition (pp. 245–254).

Falzon, L. (2006). Using Bayesian network analysis to support centre of gravity analysis in military planning. *European Journal of operational research*, *170*(2), 629–643.

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 2053951719860542.

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

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.

Fitelson, B. (1999). The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66*, S362–S378.

Fitelson, B., & Hitchcock, C. (2011). Probabilistic measures of causal strength. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 600–627). Oxford University Press. Oxford, UK.

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.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2-3), 131–163.

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.

Glymour, C. (2014). Probability and the explanatory virtues. *British Journal for the Philosophy of Science*, *66*(3), 591–604.

Goodman, B., & Flaxman, S. (2016). EU regulations on algorithmic decision-making and a "right to explanation". In *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv.org/abs/1606.08813 v1.*

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

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine*, *40*(2), 44–58.

Hahn, U. (2011). The problem of circularity in evidence, argument, and explanation. *Perspectives on Psychological Science*, *6*(2), 172–182.

Hahn, U. (2020). Argument quality in real world argumentation. *Trends in Cognitive Sciences*.

Hahn, U., Harris, A. J., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337–367.

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

Hahn, U., & Hornikx, J. (2016). A normative framework for argument quality: Argumentation schemes with a Bayesian foundation. *Synthese*, *193*(6), 1833–1873.

Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese*, *152*(2), 207–236.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological review*, *114*(3), 704.

Hahn, U., & Oaksford, M. (2012). Rational argument. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford library of psychology. The Oxford handbook of thinking and reasoning* (pp. 277–298). Oxford: Oxford University Press.

Hahn, U., Oaksford, M., & Corner, A. (2005). Circular arguments, begging the question and the formalization of argument strength. In *Proceedings of AMKLC '05, international symposium on adaptive models of knowledge, language and cognition* (pp. 34–40).

Hahn, U., Oaksford, M., & Harris, A. J. (2013). Testimony and argument: A bayesian perspective. In *Bayesian argumentation* (pp. 15–38). Springer.

Haigh, M., Wood, J. S., & Stewart, A. J. (2016). Slippery slope arguments imply opposition to change. *Memory & cognition*, *44*(5), 819–836.

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

Hájek, A., & Hartmann, S. (2010). Bayesian epistemology. In J. Dancy, E. Sosa, &

M. Steup (Eds.), *A companion to epistemology* (pp. 93–105). Oxford: Wiley-Blackwell.

Hall, S., Ali, N., Chater, N., & Oaksford, M. (2016). Discounting and augmentation in causal conditional reasoning: causal models or shallow encoding? *PloS one*, *11*(12), e0167741.

Halley, E. (1752). *Astronomical tables with precepts: Both in english and latin, for computing places of the sun, moon, planets, and comets*. W. Innys.

Halpern, J. Y., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science*, *56*(4), 843–887.

Halpern, J. Y., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science*, *56*(4), 889–911.

Hamblin, C. L. (1970). *Fallacies*. Methuen.

Hansen, H. (2015). Fallacies. In *The stanford encyclopedia of philosophy.*

Harman, G. (1965). The inference to the best explanation. *The philosophical review*, *74*(1), 88–95.

Harman, G. (1967). Detachment, probability, and maximum likelihood. *Nous*, 401–411.

Harradon, M., Druce, J., & Ruttenberg, B. (2018). Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*.

Harris, A. J., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6),

1571.

Harris, A. J., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A bayesian formalisation of faint praise. *Thinking & Reasoning*, *19*(3-4), 414–429.

Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science*, *40*(6), 1496–1533.

Harris, A. J., Hsu, A. S., & Madsen, J. K. (2012). Because hitler did it! quantitative tests of bayesian argumentation using ad hominem. *Thinking & Reasoning*, *18*(3), 311–343.

Hartmann, S., & Sprenger, J. (2011). Bayesian epistemology. In S. Bernecker & D. Pritchard (Eds.), *The Routledge companion to epistemology* (pp. 609–620). New York, NY and London: Routledge.

Hayes, B., & Shah, J. A. (2017). Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 303–312).

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition*, *133*(3), 611–620.

Hempel, C. G. (1965). *Aspects of scientific explanation*. Free Press: New York.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, *15*(2), 135–175.

Henderson, L. (2013). Bayesianism and inference to the best explanation. *The British Journal for the Philosophy of Science*, *65*(4), 687–715.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative

filtering recommendations. In *Proceedings of the 2000 acm conference on computer supported cooperative work* (pp. 241–250).

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65.

Hoeken, H., Timmers, R., & Schellens, P. J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking & reasoning*, *18*(3), 394–416.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1–55.

Hsu, A. S., Horng, A., Griffiths, T. L., & Chater, N. (2017). When absence of evidence is evidence of absence: Rational inferences from absent data. *Cognitive science*, *41*, 1155–1167.

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

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166.

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

Johnson, S., Jin, A., & Keil, F. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Johnson, S., Johnston, A., Toig, A., & Keil, F. (2014). Explanatory scope informs causal strength inferences. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one

random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601-625.

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

Keil, F. C. (2006). Explanation and understanding. *Annu. Rev. Psychol.*, *57*, 227–254.

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), 195–210.

Kirfel, L., Icard, T., & Gerstenberg, T. (2020). Inference from explanation. *PsyArXiv*.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological bulletin*, *110*(3), 499.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 461.

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence*. CRC press.

Krauss, D. A., & Sales, B. D. (2001). The effects of clinical and scientific expert testimony on juror decision making in capital sentencing. *Psychology, Public Policy, and Law*, *7*(2), 267.

Lacave, C., & Díez, F. J. (2002). A review of explanation methods for Bayesian

networks. *The Knowledge Engineering Review*, *17*(2), 107–127.

Lagnado, D. (1994). *The psychology of explanation: A Bayesian approach*. Unpublished Masters thesis. Schools of Psychology and Computer Science, University of Birmingham, UK.

Lagnado, D., Fenton, N., & Neil, M. (2013). Legal idioms: A framework for evidential reasoning. *Argument & Computation*, *4*(1), 46–63.

Laskey, K. B., & Mahoney, S. M. (1997). Network fragments: Representing knowledge for constructing probabilistic models. In *Proceedings of the thirteenth conference on uncertainty in artificial intelligence* (pp. 334–341).

Liefgreen, A., & Tešić, M. (in press). Explaining away and the propensity interpretation of probability: The case of unequal priors. *European Conference of Argumentation Proceedings*.

Liefgreen, A., Tešić, M., & Lagnado, D. (2018). Explaining away: Significance of priors, diagnostic reasoning, and structural complexity. In T.T.Rogers, M.Rau, X.Zhu, & C.W.Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 2047–2052). Cognitive Science Society.

Lippmann, R., Ingols, K., Scott, C., Piwowarski, K., Kratkiewicz, K., Artz, M., & Cunningham, R. (2006). Validating and restoring defense in depth using attack graphs. In *Milcom 2006-2006 ieee military communications conference* (pp. 1–10).

Lipton, P. (2003). *Inference to the best explanation*. Routledge.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive*

psychology, *55*(3), 232–257.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford library of psychology. The Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford: Oxford University Press.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. *Oxford handbook of causal reasoning*, 415–432.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).

Mackie, J. L. (1965). Causes and conditions. *American philosophical quarterly*, *2*(4), 245–264.

Maio, G. R., Haddock, G., & Verplanken, B. (2018). *The psychology of attitudes and attitude change*. Sage Publications Limited.

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

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, *15*(1), 75–80.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & cognition*, *37*(3), 249–264.

Meder, B., & Mayrhofer, R. (2017a). Diagnostic causal reasoning with verbal information. *Cognitive psychology*, *96*, 54–84.

Meder, B., & Mayrhofer, R. (2017b). 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.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, *58*(3), 401–415.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, *34*(2), 57–74.

Merdes, C., Von Sydow, M., & Hahn, U. (2020). Formal models of source reliability. *Synthese*, 1–29.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

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.

Moulin, B., Irandoust, H., Bélanger, M., & Desbordes, G. (2002). Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review*, *17*(3), 169–222.

Nance, D. A., & Morris, S. B. (2002). An empirical assessment of presentation formats for trace evidence with a relatively large and quantifiable random

match probability. *Jurimetrics*, *42*, 403–448.

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

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841–848).

Nicholson, A. E., Korb, K. B., Nyberg, E. P., Wybrow, M., Zukerman, I., Mascaro, S., … Lagnado, D. (2020). BARD: A structured technique for group elicitation of Bayesian networks to support analytic reasoning. *arXiv preprint arXiv:2003.01207*.

Nielsen, U., Pellet, J.-P., & Elisseeff, A. (2008). Explanation trees for causal Bayesian networks. *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, 427–434.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual review of psychology*, *71*.

Okasha, S. (2000). Van Fraassen's critique of inference to the best explanation. *Studies in History and Philosophy of Science*, *31*(4), 691–710.

Olsson, E. J., & Vallinder, A. (2013). Norms of assertion and communication in social networks. *Synthese*, *190*(13), 2557–2571.

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. (2013). Evaluating computational models of explanation using human judgments. In *Proceedings of the 29th conference on uncertainty in artificial intelligenc.*

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible*

inference. San Francisco, CA: Morgan Kauffman.

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

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of personality and social psychology*, *62*(2), 189.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision making. *Cognition*, *49*(1-2), 123–163.

Pernkopf, F., & Bilmes, J. (2005). Discriminative versus generative parameter and structure learning of Bayesian network classifiers. In *Proceedings of the 22nd international conference on machine learning* (pp. 657–664).

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

Phillips, K., Hahn, U., & Pilditch, T. D. (2018). Evaluating testimony from multiple witnesses: Single cue satisficing or integration? In *Proceedings of the 40th annual conference of the cognitive science society.*

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

Pilditch, T. D., Hahn, U., & Lagnado, D. (2018). Integrating dependent evidence: Naïve reasoning in the face of complexity. In *Proceedings of the 40th annual conference of the cognitive science society.*

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

Psillos, S. (2005). *Scientific realism: How science tracks truth*. Routledge.

Ramsey, F. P. (2016). Truth and probability. In *Readings in formal epistemology* (pp. 21–45). Springer.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*(3), 429.

Rebitschek, F. G., Bocklisch, F., Scholz, A., Krems, J. F., & Jahn, G. (2015). Biased processing of ambiguous symptoms favors the initially leading hypothesis in sequential diagnostic reasoning. *Experimental psychology*, *62*(5), 287–305.

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.

Reichenbach, H. (1953/1991). *The direction of time* (Vol. 65). Univ of California Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Rieger, L., Chormai, P., Montavon, G., Hansen, L. K., & Müller, K.-R. (2018). Structuring neural networks for more explainable predictions. In *Explainable and interpretable models in computer vision and machine learning* (pp. 115–131). Springer.

Rohekar, R. Y., Nisimov, S., Gurwicz, Y., Koren, G., & Novik, G. (2018). Constructing deep neural networks by Bayesian network structure learning. In *Advances in neural information processing systems* (pp. 3047–3058).

Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., & Tirri, H. (2005). On discriminative Bayesian network classifiers and logistic regression. *Machine Learning*, *59*(3), 267–296.

Ross, L. D., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology*, *35*(11), 817.

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.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Salmon, W. C. (1992). Scientific explanation. In M. H. Salmon et al. (Eds.), *Introduction to the philosophy of science* (p. 7-41). Englewood Cliffs, New Jersey, Prentice-Hall.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelli-

gence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, *78*(1), 105–127.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, *110*(1), 101.

Scott, A., Clancey, W. J., Davis, R., & Shortliffe, E. H. (1977). Explanation capabilities of knowledge-based production systems. *American Journal of Computational Linguistics*, 338–362.

Sherman, S. J., Zehner, K. S., Johnson, J., & Hirt, E. R. (1983). Social explanation: The role of timing, set, and recall on subjective likelihood estimates. *Journal of Personality and Social Psychology*, *44*(6), 1127.

Shimony, S. E. (1991). Explanation, irrelevance and statistical independence. In *Proceedings of the ninth national conference on artificial intelligence-volume 1* (pp. 482–487).

Singmann, H., & Kellen, D. (2019). An introduction to linear mixed modeling in experimental psychology. In *New methods in cognitive psychology* (p. 4–31). Psychology Press.

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. In *Chi'02 extended abstracts on human factors in computing systems* (pp. 830–831).

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition*, *52*(1), 1–21.

Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, *24*(2), 109–

143.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford University Press.

Sun, Y., Chockler, H., Huang, X., & Kroening, D. (2019). Explaining deep neural networks using spectrum-based fault localization. *arXiv preprint arXiv:1908.02374*.

Sussman, A. B., & Oppenheimer, D. M. (2011). A causal model theory of judgment. In *Proceedings of the cognitive science society* (Vol. 33).

Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2009). Moviexplain: A recommender system with explanations. In *Proceedings of the third acm conference on recommender systems* (pp. 317–320).

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*(1), 107–119.

Tešić, M., & Hahn, U. (2019). Sequential diagnostic reasoning with independent causes. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41th annual conference of the cognitive science society* (pp. 2947–2953). Cognitive Science Society.

Tešić, M., & Hahn, U. (in press). Explanation in AI systems. In S. Muggleton & N. Chater (Eds.), *Human-like machine intelligence.* Clarendon Press. Oxford, UK.

Tešić, M., Liefgreen, A., & Lagnado, D. (2020). The propensity interpretation of probability and diagnostic split in explaining away. *Cognitive Psychology*, *121*, 101293.

Thagard, P. (1978). The best explanation: Criteria for theory choice. *The journal of philosophy*, *75*(2), 76–92.

Thagard, P. (1989). Explanatory coherence. *Behavioral and brain sciences*, *12*(3), 435–467.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38.

Toulmin, S. E. (1958/2003). *The uses of argument*. Cambridge university press.

Trout, J. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*(2), 212–233.

Trout, J. (2008). Seduction without cause: Uncovering explanatory neurophilia. *Trends in cognitive sciences*, *12*(8), 281–282.

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.

Van Eemeren, F. H., Grootendorst, R., Johnson, R. H., Plantin, C., & Willard, C. A. (2013). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge.

Van Fraassen, B. C. (1977). The pragmatics of explanation. *American Philosophical Quarterly*, *14*(2), 143–150.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Vineberg, S. (2016). Dutch book arguments. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/dutch-book/.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review*, *10*(1), 43–62.

Walton, D. (2004a). A new dialectical theory of explanation. *Philosophical Explorations*, *7*(1), 71–89.

Walton, D. (2004b). *Relevance in argumentation*. Routledge.

Walton, D. (2007). *Witness testimony evidence: Argumentation and the law*. Cambridge University Press.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Wang, H., & Yeung, D.-Y. (2016). Towards Bayesian deep learning: A survey. *arXiv preprint arXiv:1604.01662*.

Watson, J., Whiting, P. F., & Brush, J. E. (2020). Interpreting a covid-19 test result. *Bmj*, *369*.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of

probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 781.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of cognitive neuroscience*, *20*(3), 470–477.

Weisberg, D. S., Taylor, J. C., & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision making*, *10*(5), 429.

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

Wick, M. R., & Thompson, W. B. (1992). Reconstructive expert system explanation. *Artificial Intelligence*, *54*(1-2), 33–70.

Wiegerinck, W., Burgers, W., & Kappen, B. (2013). Bayesian networks, introduction and practical applications. In *Handbook on neural information processing* (pp. 401–431). Springer.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive psychology*, *66*(1), 55–84.

Williamson, T. (2016). Abductive philosophy. In *The philosophical forum* (Vol. 47, pp. 263–280).

Wintle, B. C., Fraser, H., Wills, B. C., Nicholson, A. E., & Fidler, F. (2019). Verbal probabilities: Very likely to be somewhat more confusing than numbers. *Plos one*, *14*(4), e0213522.

Wojtowicz, Z., & DeDeo, S. (2020). From probability to consilience: How explanatory values implement bayesian reasoning. *Trends in Cognitive Sciences*.

Woodward, J. (2017). Scientific explanation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2017 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/.

Xie, P., Li, J. H., Ou, X., Liu, P., & Levy, R. (2010). Using Bayesian networks for cyber security analysis. In *2010 ieee/ifip international conference on dependable systems & networks (dsn)* (pp. 211–220).

Yap, G.-E., Tan, A.-H., & Pang, H.-H. (2008). Explaining inferences in Bayesian networks. *Applied Intelligence*, *29*(3), 263–278.

Yuan, C., Lim, H., & Lu, T.-C. (2011). Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, *42*, 309–352.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. (2017). Evaluating everyday explanations. *Psychonomic bulletin & review*, *24*(5), 1488–1500.

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, *14*(1), 1–101.

Zukerman, I., McConachy, R., & Korb, K. B. (1998). Bayesian reasoning in an abductive mechanism for argument generation and analysis. In *AAAI/IAAI* (pp. 833–838).

Zukerman, I., McConachy, R., Korb, K. B., & Pickett, D. (1999). Exploratory interaction with a Bayesian argumentation system. In *IJCAI* (pp. 1294–1299).