Discounting and augmenting in causal conditional reasoning and the influence of how judgements are elicited

Discounting and Augmenting in Causal Conditional Reasoning and the Influence of How Judgements are Elicited

Simon Hall

Doctor of Philosophy

Birkbeck College, University of London, 2021

Declaration

I confirm that the work submitted in this thesis is my own, and I confirm that the thesis references information derived from other sources.

Simon Hall

Acknowledgments

I should like to thank Mike Oaksford and Ulrike Hahn for supervision and assistance.

I also want to express my gratitude to the other academic staff at Birkbeck who have taught and helped me during my studies here.

I should like to thank all my fellow students who have studied with me through the years, and the students and assistants in the Merlin lab.

I should like to particularly say how grateful I am to 唐鳳珠, 楊韻璇, 杉田先輩, and Julius Grützke for important advice, and assistance, with my studies and beyond. I could not have started this PhD without their help.

Abstract

The research described here examines reasoning about causal relationships expressed as pairs of conditional sentences of two types. The pairs either shared a cause, and gave two effects of that cause, or shared an effect, and gave two causes of that effect.

Two phenomena that should when the causal meaning of such conditionals is taken account of are discounting and explaining away. If an effect has two causes, and the effect is known to have occurred, and then one of the causes is found to have also occurred, the probability of the other cause will fall. It is no longer needed to explain the effect.

For conditionals sharing a cause, learning that one effect is true will make the other more probable, in the case that the state of the effect is not known. One effect makes the cause more probable, which in turn makes the other effect more probable. If the cause it itself known to be true, learning one effect does not make the other more probable. The cause is all that is needed to explain the other effect.

Previous research has suggested that discounting and augmenting, which are effects of causal conditional relations, occur differently depending on whether reasoners are asked to give probabilities for one cause or one effect of a pari before and after learning that the other one has occurred, or if they are simply asked to describe the direction of the change in probability that has taken place.

This response mode discrepancy is here replicated and confirmed to be a robust effect in a series of experiments conducted over the internet.

Various data collected failed to provide a satisfactory explanation for this anomaly. A speculative explanation is described, but further research is needed.

Contents

List of tables

List of figures

Chapter 1 Introduction

1.1 An overview of this research

This thesis reports experimental research into reasoning about conditional sentences. A conditional statement joins two statements, typically by putting 'if' in front of one of them, and makes a claim that the truth of clause without the 'if' depends upon (is 'conditional' upon) the truth of the other clause.

For example: if you invest your savings in Bitcoin, you'll become rich. Conditional sentences can be represented generically by symbols, for example 'if p, then q'. Central to this research is the claim that 'if p, then q' is often lacking in important information about causality, which is found in the real world, or more precisely, found in people's understanding of the real world. In logic, the symbols can be manipulated according to the appropriate rules, and the causal direction of the relationship between the conditional clauses need not be taken into account.

A simple way to see if people do or do not use causal information when reasoning conditionally is to take pairs of conditionals which share a 'q' clause (the consequent).

For example:

If John gets a promotion, he'll smirk annoyingly

If John wins a lottery prize, he'll smirk annoyingly

These are a pair of 'common-effect' conditionals, to use a causal term. The three variables can be represented graphically as three nodes of a 'causal network', with arrows showing the causal connections. The reason that a pair of conditionals like this can show whether someone is reasoning causally is that if they are, their reasoning should show a pattern called 'discounting' or 'explaining away'. If I believe these conditionals, and I see John smirking, I

will think it more likely that he has got a promotion or won the lottery. My degree of belief in both will go up, based upon the causal relationships, and the smirk. But, if I now find out John has won the lottery, my degree of belief that he got a promotion will go back down. This pattern is an indicator of reasoning which is causal, not merely logical. If one cause is present, we need the other cause less as an explanation of the effect.

There is a mirror image of this phenomenon if we take pairs of conditionals also sharing a consequent, but which are 'common-cause' rather than 'common-effect'. For example,

If John's coughing, he has Covid

If John's sneezing, he has Covid

In this case, hearing John cough makes it more probable he will also sneeze, and vice-versa - but only if we don't know if John has Covid. If we do, we already expect sneezing and coughing, and confirming one has occurred will not make the other more likely. The effect when we don't know the status of the consequent, is 'augmenting', and is also an indication of causal reasoning.

This research aimed at replicating earlier studies (Ali, Schlottman, Shaw, Chater, and Oaksford, 2010; Ali, Chater, and Oaksford, 2011), which found discounting, and augmenting. Beyond that, it aimed at replicating an unexpected discrepancy that those studies found - the prevalence of discounting and of augmenting was different when participants were asked to give probabilities as a number, and when they were asked to describe changes in their beliefs as a direction of change only. Furthermore, this discrepancy itself showed a discrepancy between Ali et al., (2010), using child participants, and Ali et al., (2011), using adults, with respect to augmenting for common-cause conditionals.

The thesis will describe the most suitable formal representation for these conditional pairs - graphical networks called 'Causal Bayes Nets', which are used widely outside of psychology

for representing causal relationships, and calculating their probabilities. They may or may not also be related to how the mind models such relationships. At the same time, the term Bayesian applies to the preferable, and more up-to-date, statistical analyses carried out on the data. Why such analyses are preferable will be explained below. Ali et al., (2011), used more traditional, non-Bayesian, analyses. A subsequent study, also reported here, Hall, Ali, Chater, and Oaksford, (2016), began the transition towards Bayesian analysis. Hall et al., (2016), al also tested an explanation for the 'response mode discrepancy', just mentioned, where the two modes are delta (quantitative) and change (qualitative), but the explanation was not supported.

The replication was successful. Two experiments were reported previously in Hall et al., (2016), and again here. 6 more experiments are newly reported here. Discounting and augmenting were again found, confirming that reasoners take account of causal information in reasoning about conditional sentences. At the same time, the pattern of responses fell short of what a causal account would predict – the participants did not reason normatively. In addition, the response mode discrepancy was confirmed. The pattern of results in Ali et al., (2010), and Hall et al., (2016), was replicated.

Finally, an additional experiment, which diverged from the materials used for replication, was carried out with the aim of testing whether the way the reasoning scenarios were presented affected the response mode discrepancy. This did not succeed in affecting the results.

The research here also went beyond a replication of Ali et al., (2010), and Ali et al., (2011), by collecting data to examine if individual differences in memory, participants' confidence in their judgements, and how long it took participants to make those judgements, as well as how much participants felt that the pairs of causes or of effects, were correlated. Not all of the information was successfully collected, and in the end no striking new insights into the response mode discrepancy were found.

Errors in question wording meant that not all the exploratory analyses could be carried out. Overall, the replication was successful, but the discrepancy remains unexplained. Examples are given in the discussion, in the light of proposals for how Causal Bayes Nets may throw light on human reasoning, of how specific probabilities can produce unexpected patterns of results.

In the light of an independent study, Tešić, Liefgreen, and Lagnado, (2020), the discussion looks at how the specifics of natural language conditionals (here in English) may affect how reasoners see causal relationships. Since the main aim of the present research was to replicate Ali et al., (2010), and Ali et al., (2011), the materials used were based upon the materials used in those studies. The sentences used here were limited and not entirely natural to a native speaker – they are strange, rather than incorrect. For causal probabilistic reasoning, what would seem as unimportant nuances from a logical viewpoint may be of considerable importance, and more attention should be paid to syntax and pragmatics in future research.

1.2 Conditional statements: an overview

"It's very easy to slip over on those paving stones unless it's dry."

or, equivalently:

"If those paving stones are wet, it's easy to slip over on them."

Conditional statements suggest a process of conditional reasoning to a hearer. With such a statement a person can move from information about one clause to information about the other. Alternatively, a hearer may move into a discussion of the conditional itself, by replying, for example, "No, I don't think so", which suggests that just as the constituent clauses can be true or false, so can the conditional statement itself.

Being able to infer that something is so in the world because of what we know about something else is key to navigating life's snakes and ladders, and conditional statements contain the sort of knowledge we need to do that. Other types of inferences can be recast as conditionals. What conditional statements mean, and how we use them, are important questions in philosophy, psychology, and linguistics (Edgington, 1995; Nickerson, 2015; Oaksford & Chater, 2007).

It has been claimed that conditional constructions can be found in all known languages (Comrie, 1986). On the other hand, some languages are said to lack specific conditional constructions, and to require their users to infer a conditional relationship from pragmatics and the proximity of independent clauses (a process of parataxis; von Fintel, 2011). Similarly, in English, we can imply conditionality without stating it, for example, using a single independent clause: "Personally, I would have pardoned Joan of Arc", inviting the hearer to supply a conditional clause such as "If I'd been King of England at the time".

These four English conditional sentences, each with an independent clause, and a dependent clause marked by 'if', will help in introducing some different types of conditional statement, numbered from zero as the first example is of what is commonly called a 'zero conditional':

0] *If it rains, the grass gets wet*

1] *If it rains tomorrow, I'll be surprised*

2] *If it snowed tomorrow, I'd be surprised*

3] *If it had snowed yesterday, I'd have been surprised*

One way to classify conditionals is as 'indicative' or 'subjunctive'. Neither name is self-explanatory, nor is there agreement on how to assign sentences to either type (Bennet, 2003).

Subjunctive conditionals are also called ‘counterfactual’, which may be a better description when talking about English, a language in which the subjunctive mood (e.g., ‘if he be honest’) is now rarely found. Calling sentence 3 a counterfactual seems appropriate, inasmuch as someone saying it is also asserting (and presumably believes) that it did not snow the day before. The speaker isn’t strictly denying being surprised, but if they were, it wasn’t due to a snowfall yesterday. Whether 2, or even 1, is a counterfactual conditional is not agreed upon, with psychological researchers tending to draw the line more restrictively than linguists (Iatridou, 2000; Over, 2017). Doubtless tomorrow’s weather is less certain than yesterday’s. The use of the past tense to describe a future event expresses a lower probability for the if clause in sentence 2 than does the present tense in sentence 1. In linguistic terminology such a use (and the actual subjunctive mood itself) is called ‘irrealis’. For the majority of researchers who do not count sentence 1 as a counterfactual, it is an indicative conditional. In actual use, the combinations of verb forms (0: present, present; 1: present, future, 2: past, conditional, 3: pluperfect, past conditional) are not the only ones possible, and conditional statements often have another marker than ‘if’, or can be made simply by putting two independent clauses next to each other (‘I didn’t know. I would have lent you mine’).

There is no evidence that any language lacks the means to express any particular type of conditional (Karawani, 2014). Earlier research results suggesting that Mandarin speakers were at a disadvantage in reasoning with counterfactuals (Bloom, 1981) because of the lack of a dedicated grammatical construction analogous to sentence 3 above are now usually attributed to unnatural translations from English in the stimuli (Feng & Yi, 2006; Jiang, 2019).

Linguists will usually classify conditional statements as being in the conditional mode, grouped along with imperatives and questions as 'irrealis' modes, in opposition to the 'realis' mode of indicative statements. A realis mode, indicative, statement ("This is gold") is a statement with a truth value. In the irrealis mode, questions ("How are you?") and imperatives

("Don't push!") clearly aren't true or false. However, which, if any, conditional statements have truth values, and if they do, what they are, is a topic of debate and disagreement among philosophers and linguists.

Another type of conditional is the deontic conditional, dealing with obligations and permissions, as in "If you find £10 on the street, you must hand it in at a police station". By comparing such a statement with "If today is Monday, tomorrow must be Tuesday" we can see how much the meaning of a modal verb, here 'must', can differ according to the type of conditional implied: in the second sentence, 'must' denotes certainty, which is not its meaning in the deontic conditional.

A distinction that is not critical to understanding conditionals is which grammatical connective, if any, is used to mark such a statement. Philosophers and psychologists almost always prefer to examine conditional statements where the dependent clause is marked with 'if', of the type 'if *p*, [then] *q*' (equivalent to '*q* if *p*'). In practice, English has a number of ways of marking conditional statements ('unless'...) but the advantage of 'if' is that non-if conditionals can be recast using if, while not all if-conditionals can be recast in any particular alternative form. 'If' is the most general marker of a conditional and does not imply any particular type of conditional. In English, by contrast, there is no generally applicable combination of verb forms that can be used for all types of conditional. In psychological research on conditionals the sentences used may indicate a generality which is somewhat unnatural, e.g., such as 'If it rains, the grass is wet'.

Another major distinction is between 'specific' and 'general' conditionals. In the two conditional statements "If you leave that ice lolly in the sun, it will melt" and "If you leave an ice lolly in the sun, it will melt" (equivalent to the zero conditional "If you leave ice lollies in the sun, they melt"), the change between 'that' and 'an' makes a clear difference to the meaning,

which is precisely the difference between predicting what the sun will do to a particular lolly, and stating a general law which, all things being equal, holds regardless of time or place.

Sentence 0 above is an example of what is often called a ‘zero conditional’ in English linguistics and expresses a general truth. The word ‘if’ here could be replaced by ‘whenever’, which is impossible for the other sentences. The two present tenses mark sentence 0 as a general conditional.

Bennett (2003, p.16) discusses a type of conditional that he calls ‘independent’, where the independence in question is from ‘unstated particular matters of fact’, as in Bennett’s example, ‘If the river were to rise another two feet, it would be two feet higher than it is now’. Although Bennett stresses that an independent conditional can be causal, rather than logical (or definitional), the example he gives seems rather forced, relying on developments in physics (i.e., the maximum speed of a particle) up to the present, and ruling out others in the future. Bennett (2003, p.1) describes independent conditionals as ‘uninteresting’, and those that are not causal will also be uninteresting from the viewpoint of the research reported here.

A general relationship between the sun and melting ice lollies is a causal relationship, and the topic of the present research is what people take causal conditional statements to mean, and what conclusions people draw using them. The first task below is to look at ways of interpreting conditional statements that ignore causality and consider whether those ways fall short of what is needed.

Chapter 2 Conditionals in mental logic and mental models theory

2.1 The logical account of conditional inference and meaning

Two basic concerns are which inferences from a conditional are justified, and, assuming conditionals sentences have truth values, which conditionals are true. Traditional logic is binary. Conditional statements are true or false, according to the truth of the statements that make them up. Inferences are valid or not, regardless of their plausibility, and valid inferences are true, assuming the premises are true. In psychology and in philosophy, recent decades have seen a move away from binary reasoning, but these non-binary approaches to conditional reasoning have been developed in the light of, and in reaction to, the traditional approach. They often use its terminology and are defined in contrast to the older paradigm.

The logical analysis of conditionals was developed for the purpose of putting mathematics on a firmer footing in the nineteenth and early twentieth centuries. The historical development is set out in Sanford (1989). Traditional binary logic gives a clear and coherent analysis of conditional statements. In a sentence of the form 'If *p*, then *q*', e.g., 'If today is Monday, yesterday was Tuesday, the first clause (without 'if'), i.e. 'today is Monday, is called the antecedent, and the second clause (without 'then'), i.e. 'yesterday was Tuesday, is called the consequent. (A logical analysis of a sentence does not depend on whether the content is true or false.)

A conditional argument is produced by combining a conditional statement (the main premise) with a statement (the minor premise) about the truth of the antecedent or consequent. Four minor premises give four well-known conditional arguments. Adding 'p' to the main premise gives modus ponens (MP), adding '¬*p*' (not-*p*, i.e. *p* is false) gives denying the antecedent (DA), adding '*q*' gives affirming the consequent (AC), and adding '¬*q*' gives modus tollens (MT). The conclusions for these four argument types are, in order '*q*', '¬*q*', '*p*', and '¬*p*'.

In real-world reasoning, other factors affect how people draw inferences. MP and DA are 'forward inferences' and are easier for people to make than the 'backward inferences', but only if the conditional is presented with the 'if' clause first. DA and MT are more complex than MP and AC, because they include negations (Nickerson, 2015). The fundamental distinction in traditional logic is between valid and invalid arguments. MP and MT are valid, meaning that if the two premises are true, the conclusion ('*q*' and '¬*p*', respectively) is true without exception, i.e., regardless of any additional information. DA and AC are invalid – the truth of the premises does not make the conclusion certainly true, and making these inferences is an error. For the 'if and only-if', or bidirectional, conditional, all four arguments are valid.

Experimental results presenting these arguments to participants have confirmed that the traditional logic-based account of conditional statements does not match how people reason. Using statements referring to the real-world, but also with abstract statements (e.g. using symbols such as *p* and *q*), reasoning does not closely follow the logical pattern, where MP and MT are always valid, and DA and AC are not. Typically, reasoners make or accept DA and AC arguments more than half the time, and MT slightly more than those two. MP arguments are made and accepted almost, but not quite, always (Oaksford & Chater, 2007; Schroyens, Schaeken & d'yDewalle, 2001). The rate at which the four inferences are made can be affected by such factors as truth and negation, and reversing the order of the clauses (i.e. '*q*, if *p*') (Oaksford & Chater, 2007). None of these factors have a place in the traditional account of conditional reasoning.

According to the logical account, the entire conditional sentence also has a truth value. Further, a conditional sentence is 'truth functional', meaning that the truth of the conditional is a function of the truth of the antecedent and consequent. Specifically, a conditional sentence is true unless the antecedent is true and the consequent is false - meaning that, on a Tuesday, our example conditional, 'If today is Monday, yesterday was Tuesday, is true. By classical logic

this example conditional is true on all the days of the week except Monday. This truth-function, called the 'material conditional' is appropriate for mathematical logic, but is the Achilles' heel of the classical account of conditionals when applied to natural language conditionals, and everyday reasoning.

The material conditional account of the truth of conditional statements is due to Frege (Sanford, 1989). It can be represented clearly by using a truth table, listing the truth of the conditional for each combination of the truth of its two component clauses. Applying the material conditional to the natural language meanings and uses of conditional statements has led to considerable controversy. No consensus has been reached, but the problems revealed have given impetus to conditionals as a topic of research in recent decades. In particular, the failures of the material conditional as a guide to the truth, or acceptability, of conditional statements has caused causality to take a more central place in work on conditional reasoning (Oaksford & Chater, 2017; Oaksford & Chater, 2020).

The material conditional is a logical connective, joining two elements which themselves have truth values. As we will see later, the linguistic approach to conditionals which currently has the widest acceptance denies that ‘if’ is a connective (Kratzer, 2012), seeing it rather as a signifier that the clause it introduces should be taken as a restrictor on the applicability of the independent clause in the same statement. This is an example of denying that the material conditional can be imported from mathematical logic to represent the meaning of ‘if’, or the equivalents of ‘if’ in other languages. Other logical connectives include ‘and’, ‘not’, and ‘or’, and their meaning is typically represented in a truth table giving the truth or falsity of the compound statement for each combination of truth values (see, e.g., the Truth Table Generator at https://web.stanford.edu/class/cs103/tools/truth-table-tool/)of its components (thus, it would be strange to use the material conditional to stand for ‘if’ in ‘if you have a question, raise your

hand' because 'raise your hand' has no truth value). For example, 'a and b' is false except for the single case where both a and b are true.

A	B	A^B
F	F	F
F	T	F
T	F	F
T	T	T

The truth table for the material conditional is as follows:

A	B	If A, B
F	F	T
F	T	T
T	F	F
T	T	T

Any compound statements that have the same truth values are equivalent. Thus 'if *p*, *q*' is equivalent to '¬*p* or *q*', ¬(*p* and ¬*q*), etc. If we interpret the falsity of a conditional as requiring not merely the falsity of *p* and ¬*q*, but also the impossibility of *p* and ¬*q* we are interpreting the relationship of *p* and *q* as 'strict implication' (Manktelow, 2012).

As a description of how people reason in their everyday lives, classical logic is not tenable. It is a reasoning system that people must learn formally, and which requires care and attention to use (Manktelow, 2012; Oaksford & Chater, 2007). A near adaptation of classical logic to explain how people reason without training in logic is the 'mental logic' theory of Rips (1994). Three logical operators (and, not, if) join symbols to produce statements from which inferences

can be drawn - either forwards to a conclusion, or backwards, to premises. Forwards and backwards 'inference rules' specify allowable steps which can be taken to reach a conclusion or premises. Even when an inferential path exists, a reasoner may not possess all the possible rules, or may fail because the argument requires too many steps for their working memory.

Mental logic aims to be a much richer system than classical logic, because it allows for pragmatic rules based on those of Grice (1975). For example, in traditional logic, 'all' can be replaced by 'some', because 'all' implies 'some', but according to Grice's rules for language use, someone who uses 'some' when they know that 'all' could also be used ('some even integers are divisible by two') is breaking rules that enable successful communication. While Grice proposed these principles for conversation, mental logic applies them to the way individuals reason, restricting, for example, acceptable syllogistic arguments. This approach, by which conditionals are considered as having a core, semantic, meaning, and also as subject to pragmatic rules governing their acceptability or assertibility, has been proposed for several theories of natural language conditionals. At the same time, it has been questioned as tending to ‘tidy up’ accounts of conditional statements by relegating what is particularly troublesome and interesting to non-core pragmatics (Skovgaard-Olsen, 2016)

The traditional logic account of conditionals is binary, monotonic, and includes the material conditional. Each of these characteristics seems questionable when applied to natural language use. Natural language conditionals which refer to probabilities rather than certainties are common, e.g.: ‘If it carries on raining, the match will probably be cancelled’. A conditional statement such as ‘If you exercise strenuously for a few minutes every day, you’ll live longer than if you don’t exercise’ does not mean the utterer is committed to ‘If you exercise strenuously for a few minutes each day, and get run over by a bus next week, you’ll live longer than if you don’t exercise’. And, while according to the material conditional, conditional

statements with a false antecedent are true, the fact that Elvis is dead clearly does not make 'If Elvis is alive, Elvis is dead' true, nor acceptable or assertible.

Although traditional logic is no longer seen as a productive theory of conditional statements, more recent theories give accounts of how they differ from the binary and monotonic characteristics of logic, and how the paradoxes of the material conditional can be addressed. The topics reoccur in the following sections, where I will discuss the responses to the failure of logic to explain what people mean when they use conditional statements, to predict how people reason conditionally, or to give an account of how people should reason conditionally when understanding and interacting with the real world. The account favoured in this research is probabilistic, uncertain, and causal. First, however, in the next section, I will review an alternative theory (or theories), Mental Model Theory, which has attempted to address some of the shortcomings of the logical account of conditional reasoning.

2.2 Mental models theory and conditional reasoning

A more recent account of reasoning, which is nonetheless close to logic, is mental model theory (MMT; Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), in fact a developing series of theories, and an important approach to reasoning, including conditional reasoning, which aims to give an account of real-world reasoning behaviour that predicts some of the ways in which it falls short of classical logic. In particular, it gives an explanation of how people can reason differently depending on the amount of time, mental resources, or effort, they invest in their reasoning. MMT is an 'algorithmic' (Marr, 1982) explanation of reasoning, inasmuch as it aims to describe actual reasoning processes, not merely to give a normative description of what reasoning should aim to do. Later versions of MMT are also, in distinction to mental logic, described (Rips, 1994) as a semantic rather than symbolic explanation, working from imagined possible meanings of a statement, to be compared for truth against the world (or a model of the

world), rather than taking their elements as symbols to be manipulated by rules without regard to content. In its best-known form, MMT also accepts the material conditional.

In MMT, reasoning is the result of imagining possibilities, i.e., states of affairs compatible with a premise or premises. How many possibilities are represented affects the results of reasoning. Inferences that require more models to be represented are more difficult, and less likely to be made. In this way MMT offers explanations for many reasoning biases, that is to say, reasoning results which are not normative when judged against the standard of classical logic. A conclusion is drawn when the combined models of some premises give a new model beyond those represented for the premises. A reasoner may then try to find a model consistent with the premises, but in which the conclusion is false – if found, such a model is a counterexample, and the conclusion is rejected. Otherwise, or if no such search is made, the conclusion is accepted as true. MMT has a principle of truth, according to which reasoners produce models of what is true, avoiding models of what is false.

For example, a disjunction "A or B" has the following three models (sometimes described together as one model)

A

B

A B

where each line represents a possible state consistent with the disjunction. Central to MMT is the optional process of producing fully explicit models ('fleshing out'). This process will often produce extra possibilities. For the disjunction, the models are made more explicit by representing negative states, thus:

A ¬B

¬A B

A B

producing three models which combine to make up the disjunctive normal form of the premise: (A & ¬B) or (¬A &B) or (A & B). In both cases, following MMT's principle of truth, it is assumed that (¬A & ¬B) is not represented.

According to MMT (Johnson-Laird & Byrne, 1991), the basic representation of a conditional, if a then b, is the same as that for the conjunction:

A B

…

This basic representation is called an 'implicit model'. The ellipsis (three dots) on the second line represents the mental 'footnote' that a reasoner will make with the meaning that this model is not a complete representation.

The original MMT notation for the implicit representation of a (not bi-conditional) conditional statement allows for the use of square brackets to denote a term which will not appear again in additional models added by fleshing-out:

[A] B

The case where the antecedent and consequent are both true (conjunction) is a case in which the conditional statement is true, but it does not exhaust the models of true possibilities. A reasoner may produce a more complete ('fleshed out') representation with more models. The fullest representation comprises three models:

A B

¬A B

¬A ¬B

which cover all truth-value combinations of the antecedent and consequent, with the exception of the case where both are false, that is to say, the (single) case for which the conditional statement is itself false. MMT predicts that MT inferences will be drawn less often due to the presence of the relevant model (¬A ¬B) only in the explicit representation, not in the implicit version. Thus this 'fully explicit' set of models is the same as the truth table for the material conditional, adding to the case when the antecedent and consequent are both true, those cases where the antecedent is false, regardless of the consequent. Thus MMT, at least in its original form, accepts the material conditional, with its associated difficulties as criterion for real-world reasoning, e.g., that people tend to reject the two models by which the implicit model can be extended.

The models referred to correspond to true rows in a truth table of the premises; when a reasoner is exhaustive in his or her reasoning, the models represented will be the complete set of those true rows (Johnson-Laird, 1995).

In addition to the problems associated with the material conditional as a criterion of the truth of conditional statements, a further difficulty for the original version of MMT is that, as with logic-based accounts, it has no way to take account of semantic content, by which the real-world plausibility of inferences affects reasoners' judgements. One example of an distinction that is important for how people reason, but has no meaning from a logic-based account, is that between deontic and non-deontic conditionals. A later revision of MMT (Johnson-Laird & Byrne, 2002), provides an account extended with the concepts of possibility and semantic and pragmatic modulation. Johnson-Laird and Byrne (2002) list ten different combinations (sets) of models that can represent a conditional, depending on semantics, and the type of link between its clauses. In contrast to logic-based accounts, this version of MMT has a place for

non-monotonic reasoning, as for example when the conclusion “it lights” is not drawn from “if a match is struck properly, it lights” when a match is struck, by allowing for the disabling condition “the match is soaked in water”. Modulation can include information about which of the two clauses of a causal conditional is cause, and which effect. Under modulation, a conditional is no longer truth-functional, while a ‘basic’ (unmodulated) conditional is still truth-functional, but this has been seen as leading to an account whose basic and explicit representations are not wholly coherent (Evans, Over, & Handley, 2005).

In the versions of MMT described above, although probability can be included exogenously, by ‘tagging’ a model with a probability value, it can also be represented as part of the theory by dividing the total available probability (i.e., a value of 1 in traditional probability calculus) equally among true models. This method can be extended by representing more probable models more than once, thus increasing their associated probability. This way of accommodating probability in MMT will be explained more fully below, in the discussion of discounting and augmenting.

Finally, a newer revision of MMT (Johnson-Laird, Khemlani, & Goodwin, 2015) rejects the material conditional truth function, by representing statements as a conjunction of possibilities (rather than the disjunctive normal form discussed above). In contrast to the new paradigm approach to reasoning, for which probabilities are basic, and classical logic is a special case of reasoning with certainty (i.e. with probabilities of 0 and 1), the MMT version described by Johnson-Laird et al. (2015, p. 207) “implies that probabilities enter into the contents of reasoning only if invoked explicitly”.

Chapter 3 Probabilistic accounts and suppositionalism

3.1 The new paradigm approach to reasoning

Probabilistic accounts of conditional statements form part of the Bayesian approach to human reasoning, which has been dominant and productive in recent decades, to the extent that it is now often called simply 'the new paradigm' (Elqayam & Over, 2013; Evans, 2012).

The Bayesian approach to probability is older than the frequentist approach which was dominant for most of the twentieth century (McGrayne, 2011), but it was not until increases in computing power, and the development of computational algorithms based on sampling, rather than exhaustive calculations, that Bayesianism became widely used. As applying this approach became more practical, it also increasingly encouraged attempts to put the psychology of reasoning on a probabilistic basis, supplanting older approaches which viewed human reasoning as logical and deductive.

Formally, probabilities are values between 0 and 1 that conform to the Kolmogorov axioms (Hacking, 2001). From these axioms, a full and generally accepted mathematical system has been built up. What exactly this system is a mathematical account of is not agreed on. Three ways of understanding what probability is are that it derives from frequencies, from natural laws, or from beliefs. What is the probability that a tossed coin comes up heads? For a frequentist, it is the proportion of heads after tossing the coin a number of times, the more the better, with an infinite number of tosses giving the definitive answer. By contrast, according to propensity theory, a never-yet-tossed coin already has a specific probability of coming up heads when tossed. With sufficient knowledge of how the world works, a researcher could examine the coin and calculate the probability it would come up heads, if tossed. For someone who sees probabilities as beliefs, a probability is someone's (subjective) judgement of how likely a head is after tossing the coin in question. Someone may not be conscious of such a belief, or not be

able to express it, but we can elicit it by asking them to bet on the coin, or to say what bet a reasonable gambler would accept. Crucially, wherever their belief has come from, it is provisional (unless the probability is 0 or 1). The result of a single toss gives information that can improve a reasoner's belief. The way to update is by using Bayes' formula, but it is probability as belief which is specific to Bayesianism, rather than the formula, which is equally accepted by frequentists.

In probability theory, Bayesianism provides a way of combining new data with the prior information to update probabilities, allowing the process to be iterated indefinitely as new data become available. Applied to psychology, the key insight of the Bayesian approach is that what is known or taken as axiomatic about probability can be taken over wholesale and applied as a normative account of how people should change their beliefs as they learn. Intuitively, interacting with the world using existing beliefs, and improving those beliefs on the basis of experience, is an attractive account of why people think the way they do.

Making our beliefs comply with the axioms of probability, that is to say, making them probabilistically coherent, is normative. Someone who does not do this, and who makes bets according to their beliefs, is liable to make combinations of bets which he or she will certainly lose, known as Dutch books. Actual bets, or judgements saying which bets should be accepted by a gambler, are seen as ways to reveal a reasoner's beliefs or preferences in a numerical way. In practice, whether someone ever gambles or not, they must use their beliefs and preferences to make choices in an unavoidable 'game of life' and sets of beliefs which are not in line with the axioms of probability are, in general, inferior to those which are.

Markov Chain Monte Carlo (MCMC) algorithms, combined with increased computing power, are responsible for the upsurge in the practical use of Bayesian methods in recent decades. By comparing values of Bayes' formula,

$$p(H|E) = \frac{p(E|H)p(E)}{p(E|H)p(H) + p(E|\neg H)p(\neg H)}$$

the difficult-to-calculate denominator cancels out. A sufficiently long random walk through a sample space, performing such comparisons, will converge on the value of the formula. Similar sampling has been proposed to take place when people judge probabilities (Davis and Rehder, 2017).

An early application of Bayesianism to the psychology of reasoning was given by Oaksford and Chater (1994), who examined the well-known Wason selection task, originally devised as a test of reasoners' understanding of Popper's explanation of induction – examples conforming to a rule can never prove the rule true, while examples that do not conform can prove the rule false. Oaksford and Chater began by considering what behaviour would be adaptive in the real world, faced with similar tasks. They supported this by reference to Anderson's (1990, 1991) theory of rational analysis, which gives a process to derive adaptive reasoning and behaviour when time and cognitive resources are limited. They contended that, while only a non-white swan can provide certainty as to whether all swans are white, in practise observers can increase the information they have about swans by looking at any swan nearby, including white ones, inasmuch as they are relatively rare, and carry some information about the world when they appear. This sees the task that participants in Wason tasks are actually carrying out, not as attempting logical refutation, but instead probabilistic information gain. Predictions made based on this approach gave a much better fit to experimental data than those of binary logic.

Analysis of reasoning data informed by the new paradigm has shown that people's judgements, which fall short of the normative standards of logic, and use the restricted

knowledge relevant to logical tasks, appear much better judged by probabilistic standards, which both expect the application of world knowledge beyond that supplied by an experimenter, and also expect reasoners to be cautious about taking knowledge, and the sources of knowledge, as reliable or certain. The probabilistic approach to reasoning differs from approaches such as those of logic and MMT described above inasmuch as it does not so often see experimental results as revealing errors due to shortcomings in capacity and performance. Earlier experimental data is often re-interpreted as showing participants persisting in using reasoning strategies appropriate to real-world tasks, rather than following explicit or implicit instructions given by experimenters. Typically, human reasoning appears better, that is to say, more normative, judged according to the more complex, semantically rich, standard of the probabilistic 'new paradigm'. The central insight is that ‘everyday rationality in guiding thought and action seems to be highly successful’ (Oaksford & Chater, 2001, p. 349). The new approach has been applied productively to understanding the meaning and use of natural language conditional statements.

3.2 Probability and conditional statements

As pointed out above, natural language conditionals are not well represented by an account which is monotonic and binary. The logical account of conditionals describes monotonic reasoning. If the conditional and the antecedent are true, the consequent is true, and there is no place for further information to change that conclusion. By contrast, a Bayesian account of reasoning always allows for new information to change previous conclusions, with the exceptions of beliefs with a probability of 0 or 1, which correspond to false and true in the logical account.

Natural language conditionals are used without ruling out revision due to new information that becomes available after an inference has been drawn (Nickerson, 2015). For example, *if x*

is a bird, x can fly and *Tweety is a bird* lead to the conclusion that *Tweety can fly*, but reasoners are not unable to reach a different conclusion on finding out that *Tweety is an ostrich* (Oaksford & Chater, 2007). This kind of reasoning is also called *defeasible*.

Similarly,

if you carry on smoking like that, you'll get cancer

could be said by someone who uses English conditionals correctly, and would not be intended, or understood, as taking only the values of true or false for its clauses or the whole conditional. Imposing binary truth values onto natural language conditionals seems to give them meanings which constrain their use more narrowly than the way people use them every day. Further difficulties with the logical approach which make a probabilistic account seem preferable can be seen by looking again at the material conditional.

3.3 Suppositionality

While according to logic, if a conditional and its antecedent are true, the consequent must also be true, it seems clear that natural language conditionals are often asserted for relations are probable but not certain. The name 'normic conditionals' has been proposed for conditionals where MP is not always justified, as opposed to 'strict' conditionals which have no exceptions (Nickerson, 2015). For example, if you treat people fairly, they'll treat you fairly in return is surely to be taken as a good working rule without implying it is always true. 'If we meet again this time next year, we'll both be another year older' seems to be a strict conditional. As well as using conditionals which express probable relations, people are ready to reason from evidence which is not certain, for example making an MP inference without being 100% sure of the antecedent. It is reasonable to infer MP from 'if you're bitten by a black widow spider, seek medical assistance immediately' even if the person bitten is not an expert on spider recognition.

If people assign probabilities to conditionals, how might they obtain those probabilities? In probability theory a conditional probability is defined as the conjunction of probabilities divided by the probability conditioned on: $Pr(q|p) = Pr(q \ \& \ p) \ / \ Pr(q)$. This is often made intuitively clear by using a Venn diagram, where the area representing $Pr(p)$ is re-conceived of as an updating of the whole area, or all possibilities, and the proportion of the area for $Pr(q)$ compared to that new background is then the conditional probability. For psychological accounts, people are usually assumed not to reason in this way from base rates to conditionals.

A procedure for judging the truth of a conditional statement was famously put forward by Ramsey (1931, p. 155). The statement is brief, and its interpretation is not unambiguous: "If two people are arguing 'If *p* will *q*?' and are both in doubt as to *p*, they are adding *p* hypothetically to their stock of knowledge and arguing on that basis about *q*". In this original form, the procedure will give a binary, true or false, result, rather than a probability. A probabilistic version, by which reasoners should suppose the antecedent to be true, then judge the probability of the consequent, is central to a group of accounts which can be called 'suppositionalist'.

For most probabilistic accounts, the probability of the conditional is equated with the conditional probability, so that: $Pr(\text{if } p, \text{ then } q) = Pr(q|p)$, giving conditionals a value from 0 to 1, rather than being only either true or false. The value can be a metric of truth, of belief, of assertibility, or of acceptability. This 'probability conditional' (Adams, 1996) is commonly called simply 'the Equation' (Edgington, 1995). As an account of the judgements people actually make, it has received empirical support (e.g., Evans & Over, 2004). This probability has also been proposed as the value of the truth table cases with false antecedents (Jeffrey, 1990), giving a de Finetti truth table, where the values of the false antecedent entries are set to this conditional probability.

3.4 Supposition and counterfactuals

Probabilistic accounts also offer an intuitive approach to understanding subjunctive, or counterfactual, conditionals, and 'open' indicative conditionals, for which the status of the antecedent is unknown. The material conditional seems unable to handle counterfactuals in a meaningful way. If the antecedent is false, then by the material conditional, the conditional statement is true. Although there are different ways to define a counterfactual (Bennett, 2003; Nickerson, 2015; Over, 2017), and disagreement as to whether a distinction should be made between counterfactual and subjunctive conditionals, conditionals whose antecedent is false, or at least believed to be false by the utterer are generally taken to be counterfactuals. Conditional statements for which the utterer does not know the status of the antecedent, e.g. 'open' indicatives, may be considered to require a similar mental process to 'canonical' counterfactuals.

These cases demonstrate that using the material conditional to analyse natural language conditionals brings extra obscurity rather than clarity. A solution to this problem is to say that no conditionals (not merely special cases such as 'if you have a question, raise your hand') have truth values – this is the position, for example, of Edgington (2011). Nonetheless, if our goal is to account for how people reason with conditional statements, rejecting all truth values for conditionals also seems questionable. 'That's not true!' is surely an appropriate and easily understood response to many conditional statements, for example, 'if you buy high and sell low, you'll make a success of investing'. Conversely, 'that's right' is an acceptable response to many conditional statements.

One response has been to say that logical conditionals have truth values, and natural language conditionals do not (Stalnaker, 1968). Another approach has been to modify the conditional truth table to address the cases of true conditionals with false antecedents. As

mentioned above, de Finetti (1974) proposed that the two false antecedent entries in the table should be given neither-true-or-false values for the overall conditional, and Jeffrey (1990) proposed giving the false antecedent cases the value of the conditional probability, Pr(*q*|*p*). A different approach has been to claim there are pragmatic restrictions which make some conditionals not acceptable or not assertible, regardless of their truth value.

Grice (1989) claimed that although 'if a triangle has four sides, every day is a holiday' is indeed a true statement, if someone asserts such a true conditional when he or she knows the truth of the antecedent and / or the consequent, the assertion would not be helpful, as expected, to a conversation partner.

Jackson (1990) gave a different criterion for when a conditional is assertible, to be considered in addition to its truth. Jackson's criterion was that, taking a conditional as a general rule, rather than a description of a specific case, its assertibility requires also that a high probability of the antecedent is associated with a high conditional probability. For example, suppose that if at some point triangles start having four sides, we will know that it's likely that every day is now a holiday. But the unsatisfying nature of 'if a triangle has four sides, every day is a holiday' is surely precisely because it is an example of a conditional where the two clauses are unrelated to each, either by definition or by causality.

The Ramsey test was developed into the 'possible worlds' approach to understanding conditional reasoning (Ramsey, 1931; Stalnaker, 1968; Lewis, 1976; Kratzer, 2012). To assess 'if *p*, then *q*' by the possible worlds equivalent of the Ramsey test, one must (ideally being ignorant of the truth of p), assume p to be true, i.e., assume the truth of *p* to be added to one's stock of knowledge and consider what would be the truth value of *q* in such an imagined world. The world chosen should be as close as possible to this one.

This world may not be easy to identify. Consider a world in which JFK was not killed by Oswald in Dallas. 'If Oswald had not shot JFK, he would have won a second term' requires considering what kind of world, with that antecedent as true, most resembles the real one. Surely one in which JFK died in Dallas of a heart attack, or because of a car bomb, differs less from this one than a world in which he survived his visit. Intuitively, though, we don't think JFK would have died anyway. An understanding of counterfactual conditionals which is based on causality (rather than one which explains causality via counterfactuals) seems necessary to get the results which seem intuitive. If Oswald's shooting caused Kennedy's death in Dallas, without that cause Kennedy would not have died. Possible worlds are not needed to understand this if we see causal relations as deciding what the world is like.

Regardless of what a probabilistic account of conditionals should look like, it is surely the case that the material conditional is not a helpful description of how reasoners handle conditionals with false antecedents. According to the material conditional, all false antecedent conditionals, including counterfactuals about the past, are true. Thus 'if JFK was alive in 1965, he died in 1960' is true.

Counterfactual reasoning is generally seen as based upon causal reasoning (Edgington, 2011; Hoerl, McCormack, & Beck, 2012; Lewis, 1976). If we reason that changing the truth, or probability, of the antecedent (alone) would change the truth, or probability, of the consequent, we are taking the antecedent to be a cause of the consequent.

3.5 Causality and conditionals

Empirical results of how people reason conditionally support the claim that they take causes and effects (which have no place in traditional logic, and no central place in MMT) into account when reasoning.

Byrne, (1989), and Cummins, Lubart, Alksnis, and Rist, (1991), investigated the effects of alternative causes and disablers in conditional reasoning, in what Oaksford and Chater, (2003, 2007, 2017), call the explicit, and implicit, suppression paradigms, respectively.

For the explicit paradigm, Byrne (1989) presented participants with conditionals (e.g., 'If the key is turned, the car starts') alone, or paired with another conditional referring to either an alternative cause (e.g., 'If jump started, the car starts') or a necessary additional condition, that is to say, the negative of a disabler (e.g., 'If there is fuel, the car starts'). The alternative cause conditionals reduced ('suppressed') DA and AC. This makes sense, since an alternative cause provides a causal path via which the effect can be caused without the most salient cause (e.g., the key is turned), and so the presence of the effect need not mean that cause has already occurred. The disabler conditionals suppressed MP and MT, also as expected. A disabler means the most salient cause is no longer a reliable cause of the effect, and the absence of the effect need not imply that cause did not occur.

With the implicit paradigm, Cummins et al. (1991), looked at how people endorsed inferences based on conditionals presented alone, but which had been pretested for the number of alternative causes and disablers that participants were able to think of. Here the suppression effects occurred for the same inferences as in Byrne (1989): DA and AC for conditionals with many alternative causes, and MP and MT for conditionals with many disablers.

Though the results of these two experimental paradigms were qualitatively similar, Oaksford and Chater, (2017), compared the explicit (Byrne, 1989), and implicit, (Cummins et al., 1991), experimental results, and showed that the effects were quantitatively stronger for the explicit paradigm. The explicit paradigm supplied participants with alternative causes, or disablers. These were not obscure or surprising, and presumably participants already knew of

them. Thus it seems that, even when people have the same knowledge, in long-term memory, of causal relations in the world, more, or less, complex models can be constructed for reasoning.

By some definitions of what is a causal conditional, the relationships described in the materials used by Byrne (1989) were not causal. Rather, they referred to actors' intentions, or their personal dispositions. Oaksford and Chater (2017, p. 333) say their position is that 'the difference is [not] consequential'. In fact, in everyday language, the language of causality is applied freely to such factors. 'If her man hadn't made her jealous, Ruth Ellis would not have killed him', in which causality is central to the meaning of 'kill' and 'make'. In philosophical descriptions of causation, which have strongly influenced psychological accounts, there is a preference for simple, mechanical, relationships, such as colliding billiard balls. In everyday language, the language of causation is naturally used both for complex mechanisms with many intervening steps ('if there hadn't been a Big Bang, Sgt. Pepper would not have been released') and for the results of biological agents goals, intentions, and actions. In fact, (Edgington, 2011; Michotte, 1963) people perceive agency and causation in the simplest interactions of abstract geometric shapes. Whether this is justified in fact cannot be satisfactorily decided before a firmly grounded definition of causality is produced. It is clearly, however, a psychological truth about how people reason about the world (Oaksford & Chater, 2011).

When conditional statements that are apparently not causal, are used (as they often have been) as experimental stimuli (for example, of the type 'if this letter, then that number') the results can effectively be modelled by assuming that reasoners are seeing them as causal, in the sense that MT, DA, and AC inferences are interpreted as giving evidence about the conditional itself – the categorical premise is taken as a counterexample (Oaksford & Chater, 2013, 2017). Modelled in this way, a good fit was made to data from a large number of studies using such stimuli (Oaksford & Chater, 2013; Schroyens, Schaeken, & d'yDewalle 2001). For example, in meaningful, not abstract, terms, 'if the key is turned, the car starts' together with 'the car

didn’t start’ is not taken as suggesting the key wasn’t turned, but as an assertion that it was, and something (a disabler) stopped the car starting. Here the pragmatics of language interact with the causality of the conditional: someone who says ‘my car didn’t start’ only to confirm later that they had indeed never turned the key would be sure to annoy their conversational partner. It seems that these reasoning methods are applied, quite inappropriately, to abstract conditionals. This would be another example revealed by the new paradigm of reasoners’ persistence in handling experimental tasks with tools that only have value for real-world, meaningful, situations.

Chapter 4 Inferentialism

4.1 A true antecedent and true consequent do not guarantee a conditional's acceptability

Another insight into how people use natural language conditionals comes from considering the superficially problem-free truth table entry where both clauses are true. It would seem that such cases are definitely true, or maximally acceptable, or assertible. However, if we yoke together arbitrary, true, statements, to make a conditional, it is likely to seem unnatural:

If the Spice Girls were commercially successful, kangaroos are indigenous to Australia

The apparent cause of the strangeness of such sentences has led to the name 'missing-link conditionals' (Krzyżanowska, 2019). They are crucial to a recent new turn in the analysis of natural language conditionals, which retains the probabilistic insights of the new paradigm, but builds upon them by proposing that a more-than-probabilistic link is needed between the clauses.

The unsatisfactory nature of the material conditional truth value of false-antecedent conditional statements was noted above. The truth table entry corresponding to MP, is both true according to the material conditional, and also represents the clause values most likely to lead to endorsement, empirically, by reasoners. Nonetheless, true antecedent, true consequent conditionals reveal an apparent inadequacy in not only logical, MMT, but also suppositional theories of the meaning, or acceptability, of conditional statements. None of these accounts in their original form distinguishes 'missing-link conditionals', which have an antecedent and a consequent that lack a meaningful connection. Missing-link conditionals are often felt to be 'odd' (Douven, 2017). For example (from Krzyżanowska & Douven, 2018):

If 2 plus 2 is 4, then Paris is the capital of France [1]

From the perspective of logic, this statement is both a well-formed conditional, and true. The antecedent and the consequent are both true, and, consequently, the statement as a whole is true. But a link between arithmetic and the status of Paris cannot be brought to mind, and lacking such a link, the conditional seems inappropriate without being false. Logical accounts of conditionals have no requirement for a link, as the material conditional shows. Neither, however, do the basic MMT or suppositional accounts, though it has been proposed to extend both with additional pragmatic requirements, such that missing-link conditionals are not assertible. It seems that a natural language conditional conveys more than the corresponding statement in logic. We expect a link of some sort between the two clauses, or expect that the utterer believes there is such a link (Baratgin, Over, and Politzer, 2013; Politzer, Over, and Baratgin, 2010).

For statements such as

If you cook it for more than ten minutes, it will taste bad [2a]

If bond yields are up, Trump must be tweeting again [2b]

their greater acceptability in comparison to the example above ([1]) is not due to the clauses being more true or more probable. Rather it is because we can easily imagine some mechanism or process in real world that links the two parts of the statement. What the link is may be obscure or unknown to us, and the link may well be somewhat unreliable, but our knowledge of the makeup of the world makes it, in both cases, plausible. Even a listener who preferred versions with 'less' instead of 'more', 'down' instead of 'up' would surely not feel that the utterer didn't know how to use conditional statements. What is missing in 1 is not missing in 2a and 2b.

The material conditional produces an infinite number of true missing-link conditionals, inasmuch as any conditional with a false antecedent, and all possible consequents, is true. As

noted above, this is one of the reasons that the material conditional is so often rejected in accounts of conditional reasoning.

4.2 Is there a semantic or a pragmatic requirement for a link in conditionals?

The status of the required link is not clear (Douven, Elqayam, Singmann, van Wijnbergen-Huitink, 2018; Krzyżanowska, 2019; Oaksford & Chater, 2020; Skovgaard-Olsen, Singmann, Klauer, 2016). An unresolved question is whether the requirement for a link is pragmatic or semantic (Douven, Elqayam, Singmann, van Wijnbergen-Huitink, 2019; Krzyżanowska & Douven, 2018). Another is whether the strength of the link is probabilistic (Skovgaard-Olsen, Collins, Krzyżanowska, Hahn, & Klauer, 2019) or due to a relationship that precedes or underlies a probabilistic association (Douven et al., 2018). Skovgaard-Olsen et al., (2016) take the valence of delta-p as a measure of relevance, distinguishing delta-p >0 as positive relevance from delta-p <0, negative relevance, and delta-p =0, irrelevance. Delta-p is $Pr(q|p) - Pr(q|\neg p)$.

The semantic / pragmatic distinction is neither universally accepted, nor are the criteria for deciding between the two clear (Oaksford & Chater, 2020; Skovgaard-Olsen et al., 2019).

The standard probabilistic account of conditionals, based on the Ramsey Test and the probability conditional sees missing-link conditionals as true, or probable, or acceptable, or assertible, from a semantic viewpoint. That they seem, intuitively, wrong is seen as due to pragmatic requirements, not calling into question the suppositional approach itself (Skovgaard-Olsen et al., 2019). Such a suppositional account can be seen as resembling MMT in using pragmatics as a way to protect the essence of the theories.

Krzyżanowska (2019) distinguishes 'missing-link' conditionals, where true clauses whose content seems obviously unrelated are yoked together by 'if', from 'false-link' conditionals, where the antecedent and consequent relate to the same topic, but the link implied will be rejected as deceptive.

Thus, to investigate whether MMT and suppositional theories face a serious challenge from 'inferentialism', empirical investigations of whether the requirement for a meaningful link between the causes of a conditional is pragmatic or semantic seem to be needed. However, there is no generally accepted definition of pragmatic effects, either in linguistics or in philosophy, and thus no clear line dividing them from semantics (Birner, 2012; Oaksford & Chater, 2020).

In some cases, a pragmatic meaning seems to depend upon the particular situation in which an utterance is used, and upon the intentions of the user. The meaning of "Marmite again! Great!" is hardly understandable without knowledge of who says it, and what they think about the tasty spread. If the judgement being expressed on Marmite is negative, it is pragmatic, because the core meaning of 'great' is positive. This would be a 'conversational implicature'. A different kind of pragmatic effect is a 'conventional implicature' where the pragmatic effect is normally accessible without knowledge of the circumstances of use. Examples are often given distinguishing the connectives 'and' and 'but'. In a truth table, 'young and carefree' and 'young but carefree' will have the same entries. Thus, their difference, as to whether 'young' suggests 'not carefree' can be assigned to a 'conventional implicature'. Of course, the difference of 'but' from 'and' could also be seen as semantic and a core meaning, and the logical approach to such connectives itself seen as lacking (Bach, 1999; Birner, 2012). An account of language in which 'but' and 'and' have the same basic meaning seems to be missing something essential.

An example of a commonly accepted test for whether a meaning is pragmatic is 'cancelling'. 'Grass is green. Actually, it's not green at all' seems unacceptable, while 'Marmite is great. Actually, it's not great at all' could probably be uttered by a Marmite-hater without making it appear they were having a breakdown.

Using tests such as cancellability, Skovgaard-Olsen et al., (2019) concluded that the requirement for an inferential link between the clauses of a conditional was a conventional implicature, which they concluded was evidence for the link being both semantic and not forming part of the truth-conditional core of the meaning of 'if'. Nonetheless, conventional implicatures are, as we have seen, more closely linked to the meaning of a particular word whose meaning is diverged from, rather than to the situational intent of particular language users. Since 'if' is merely the most general version in English of a relationship between clauses that can be expressed in several ways, or even implied by putting two independent clauses next to each other, with no lexical marker, such empirical results may only apply to a specific expression in a particular language. Whether a conditional statement in English which does not use the word 'if' implies a conventional implicature is presumably an empirical question which cannot be resolved by considering the topic of discourse, or by reference to results such as those in Skovgaard-Olsen et al. (2019). Similarly, the status of missing-link conditionals in other languages than English constructed using words analogous to 'if' (such as 'wenn', 'si', 'rúguǒ' in German, French, Mandarin) cannot be predicted from that of English 'if' if the requirement for a link is a conventional implicature. Thus this result, if accepted, would seem to lead to a fragmentation of the status of conditional statements by language and expression.

4.3 Causal links

Clearly the inferential account of conditionals is a work in progress, and we might expect not a replacement of suppositionalism by inferentialism, in the way the new paradigm itself has to considerable extent replaced classical logic in reasoning research. Rather, as Oaksford and Chater (2020) suggest, the use of the CBN formalism, already productive within the new paradigm / suppositional framework (Oaksford & Chater, 2017), is likely to be fruitful for implementing inferential models of reasoning. The default link will in such an account be

causal, and its expression may be in a deductive, inductive, or abductive conditional (Oaksford & Chater, 2020; Skovgaard-Olsen, 2019).

Although CBNs have helped our understanding of how to formalise cause and effect relationships to move forward (Pearl, 2000, 2014; Pearl & Mackenzie, 2018; Sloman, 2005), such applications typically require the use of causal understanding to move beyond associational data: ‘probabilistic causality cannot be regarded as a program for extracting causal relations from temporal/probabilistic information’ (Pearl, 2000, p. 252).

Causation is far from a clear concept. As Hume pointed out, causation is not observable, merely inferable, but if we infer a cause-effect relationship from sensory data, we seem to be assuming that what happens in the real world causes our sensory experiences.

Towards the beginning of the last century, causality was sometimes seen as an unnecessary concept for science.

Nonetheless, causation remains an essential part of everyday thinking (and scientific explanation) and, even if not clearly understood, it cannot be successfully avoided in a psychological account of reasoning. It is also not only a common topic of discourse ('make', 'force', 'cause') but encoded in the meaning of many transitive verbs, e.g. 'kill' (De Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017; Wolff, 2003).

A difference between causation in philosophical accounts, and in everyday reasoning, is the tendency of the former to concentrate on mechanistic accounts analogous to colliding billiard balls, while the latter is ready to include biological agents and complex situations as causes. 'Guns don't kill people, people kill people' expresses rich causal reasoning that is mostly not seen in research into causal conditionals, nor are people's conscious or unconscious choices seen as variables suitable for the nodes of a CBN (see next section). Oaksford and Chater (2020)

refer to the work of Barwise and Perry (1983) as providing an account of causation which more closely resembles that of its everyday use outside of the laboratory or AI lab.

Here it is worth considering how far research on conditionals has tended to reflect the preferences of philosophers (and psychologists) for understanding over action. For example, ‘Conditional sentences are used to express our methodological dispositions: I express my disposition to come to believe that the plumber has arrived by saying "if the doorbell rings, that will be the plumber."’ (Stalnaker, 1968, p. 103).

As Pearl (2000, p. 253), 'the bulk of our causal knowledge' comes from 'explicit causal sentences', which is usually qualitative rather than quantitative, i.e. lacking information about probabilities or correlations. A specific causal conditional statement can be a link in a long chain of causal utterances, including statements from people whose identity is no longer known, and people no longer alive, but the beginning of such a chain must be based on observation of the world. On the one hand, this shows that even if our minds work with beliefs in a probabilistic fashion, that information must be encoded and decoded into the discrete, symbolic, framework of language as it passes between individuals. On the other hand, it is worth thinking how much of that knowledge is not primarily intended for inferring the state of the world from limited knowledge. Rather, it is often intended to give a policy for updating the world itself in line with our preferences (or avoiding an unwelcome change, or update). If Alice says to John, 'if you bang the vending machine just at this spot here, you'll get a free chocolate bar', she is probably not trying to help him to infer whether there is a chocolate bar in the tray from knowledge of recent banging, without the need to look, or to reason that there was earlier banging upon seeing a chocolate bar. She is helping him with a rule for changing the world, and belief updating may well next occur when the machine runs out of chocolate: no bar means there's no point in banging any more, although John may try a few times to see if the status of the rule has changed, or if it was merely uncertain to begin with. Nonetheless, a conditional

supplied and used as a rule for making desirable changes in the world, or avoiding undesirable changes, is not a special type of conditional, and can be used for inferring effects from causes, or causes from effects. John could decide to use the chocolate bar conditional in this way, as a rule for updating his beliefs.

As a rule for action, a conditional *must* be causal. Thus causality is the only foundation on which these two uses of conditionals, intervening in the world, and updating beliefs, can be unified.

Chapter 5 Causal Bayes Nets

5.1 Overview

Causal Bayes Nets are graphical representations of causal relationships in the world. Qualitatively, CBNs make it easier to grasp which causes contribute to which effects. Most importantly, a CBN shows which variables give no information about the value of some other variable, according to what else is known about the state of the network. Quantitatively, CBNs have been associated with the development of algorithms making the exact calculation, or alternatively the approximation, of the states of a system of causes and effects, computationally tractable. A CBN shows which causal influences can be ignored, and which need to be taken into consideration. Algorithms also exist for inferring the parameters and structure of a CBN from data, although learning causal relations without the use of a previously specified model is considered impossible: "causal questions can never be answered from data alone" (Pearl, 2018, p. 350).

Historically, CBNs were developed from networks showing statistical associations between variables. Causality is asymmetric, where association is not: if X and Y are associated statistically, in a non-causal network, X -> Y (X implies Y) and Y -> X have the same meaning – in fact, in a CBN, the directed arrow denotes causality. When constructing a CBN, it is necessary to choose the causal direction of a relationship between variables. If, for example, observation of the real world shows that Y always precedes X in time, X cannot be the cause of Y, and a CBN will show Y -> X. This adaptation of Bayesian networks to take account of causality was the result of research in the 1990s (e.g. Pearl, 2000, 2018; Spirtes, Glymour, & Scheines, 1993).

5.2 Components of a CBN

A CBN is a directed acyclic graph, with nodes (alternatively called vertices) standing for variables, and edges standing for causal connections.

The graph is directed because causal connections are directional, from cause to effect. It is acyclic because the computational algorithms used for inference on a CBN require that two nodes are not both causes and effects of each other. Mutual causation, as in a feedback loop, can be handled by adding a temporal dimension to the graph, and assessing the causal influences between the nodes in alternate directions at successive time points (Korb & Nicholson, 2011).

In practice, we can consider absent edges as the third, and defining, component of CBNs, along with nodes and directed edges. When two variables are not connected by a causal arrow, the graph is asserting that they have no direct causal influence on each other. Any influence between the two must be indirect, via pathways through other variables. The ability to reveal when the state of one variable gives no information about the state of another variable is the central benefit of using CBNs. It allows causal systems to be understood more easily and intuitively by looking at their graphical representation. It also makes calculations about the states of a causal network tractable.

There are four causal explanations for a statistical association between two variables, x and y, in a CBN. x may cause y, y may cause x, x and y may share a common cause, or x and y may share a common effect (or a combination of more than one of these four). The third explanation is familiar as a source of error when we fail to consider ("condition on") the common cause. For example, if we observe a correlation between prevalence of men wearing no tops, and ice-cream vans, we might incorrectly try to explain one in terms of the other if we do not consider their common cause, summer weather. While the third explanation may lead to 'confounding', the fourth can produce 'collider bias'. If good looks and a good voice are both

advantages for a career as a pop star (i.e. they are causes of such a career, not effects), we may examine pop stars, or non-pop stars, and infer (incorrectly) evidence that good looks imply a good voice, or vice-versa, in the population in general. Simple graphical structures of the common cause or common effect type are central to the present research.

In a CBN a path is one or more edges connecting two variables, and a directed path is a path for which all arrows in the path are in the same direction. That is to say, a directed path includes no colliders, at which a variable is the effect of more than one cause. If there is no directed path between two variables in a CBN, the CBN is stating that neither variable affects the other variable.

The nodes (variables) in a CBN are parents or children of nodes of nodes to which they are connected directly, and descendants or ancestors of nodes to which they are connected either directly, or by a directed path. In a CBN, a node with no parents is exogenous; a node with one or more parents is endogenous.

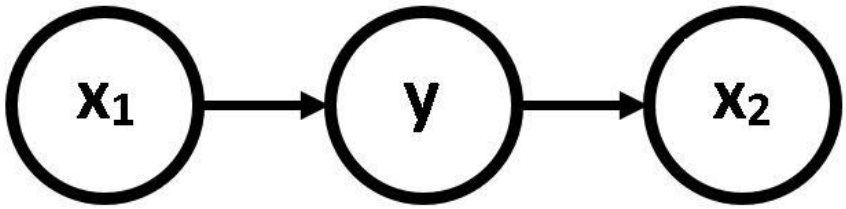
While a CBN of nodes and edges is useful in helping a qualitative understanding of the causal system it represents, a CBN typically also has parameters that can give the quantitative state of the variables in each network state. Each node has an associated conditional probability table, listing the values of the node for each combination of the values of its parent nodes. Nodes with no parents have a 'degenerate' probability table, that is to say, one that gives only the prior probability, or base rate, of that variable (Korb & Nicholson, 2011). Alternatively, a network can be parameterised with 'causal strengths', denoting how likely a cause is to produce an effect, in the absence of other causes (or how likely it is to inhibit the effect). Where more than one cause exists for an effect, an integration rule is also required, stating how the causes combine (Cheng, 1997; Pearl, 2000; Rottman & Hastie, 2014).

The simplest CBN specifying a causal system consists of two nodes, a cause and an effect. When parameterised, both such a CBN and its equivalent undirected version allow the state of the cause to be inferred from the state of the effect, and vice-versa. The CBN, unlike the undirected graph, models the effect of intervention (often called the 'do'-operation: Pearl, 2000, 2018). Intervening on the system by setting the cause to a particular value can change the value of the effect, but setting the value of the effect cannot change the value of the cause.

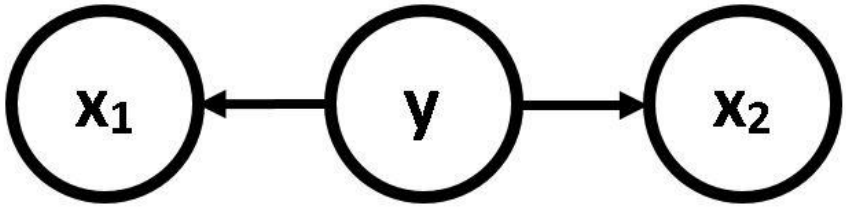
The next step up in CBN complexity is three-node networks, which will be the object of the present research. Three-node networks come in three flavours: chain, common-effect, and common-cause (see Figure 1).

Figure 1

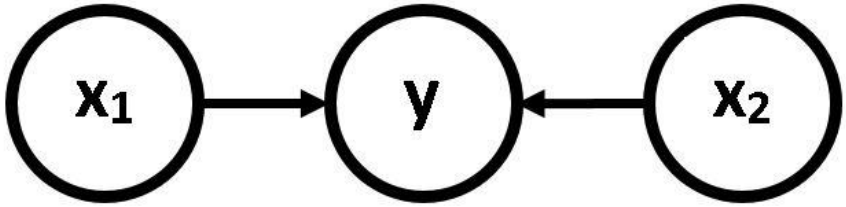
Three-node CBN Structures



1a: a causal chain network – an unconditionally open path



1b: a common-cause network – an unconditionally open path



1c: a common-effect network – an unconditionally closed path

Three-node CBN structures. In 1a and 1b, x_1 and x_2 are d-connected. In 1c, x_1 and x_2 are d-separated.

5.3 Paths and d-separation

To discover unconditional, and conditional, independencies between variables in a CBN, we must pay attention to d-separation (directed graph separation). Two nodes on a graph are d-separated if there is no open path between them; otherwise they are d-connected. A path is a route via edges between two variables, and is open (or unblocked, or active) if no collider lies on the path. If the path includes a collider, it is closed (or blocked, or inactive). A collider is

simply an effect with more than one cause, that is to say a node with more than one incoming arrow.

For example, in a three-node common-effect CBN, the two causes are d-separated by the collider, which is the effect variable. Thus this CBN asserts that the two causes are unconditionally independent of each other; they are not associated. Conversely, in a three-node common-cause CBN, the two d-connected effect nodes are d-separated when conditioning on the value of the cause.

Conditioning on a variable lying on an open path blocks the open path. Thus, in either of the two open-path arrangements of three nodes - as a chain, or as a cause with two effects - if we know the value of the node in the middle, an association between the two outer nodes is blocked. Conditioning on the intercepting variable closes the open path. Conversely, conditioning on an collider node intercepting (and thereby closing) a path opens the path. In the common-effect structure, conditioning on the collider will open the path. Thus conditioning on an intercepting node is able to flip the path which it lies on from open to closed, or from closed to open. A blocked path can also be opened by conditioning on the descendant of a collider, or some combination of the collider and descendants, but as this research is restricted to three-node networks, it will not include cases where a collider has a descendant (effect) variable.

5.4 The Markov Assumption

If we have (and this is, of course, a big if) a correct constructed CBN, accurately specifying the relationships between the variables we are interested in, then the implications of d-separation can be seen in two rules for CBNs, the Markov Assumption (MA), and faithfulness (Pearl, 1988; Verma & Pearl, 1988).

The Markov Assumption states that, if two variables X and Y (X and Y can each be a set of variables) are d-separated by conditioning on an intervening variable Z (or a set of variables, including an empty set), X and Y are statistically independent of each other, conditional on Z. In the two simple open networks in figure 1 (1a and 1b), conditioning on y makes x1 and x2 independent. Once y is known, we have all the available information about the state of either x, and the knowledge of the other x is no longer informative. This effect of conditional d-separation is also known as 'compatibility' (Glymour & Greenland, 2008).

When dealing with larger networks, inferences on the value of a variable can be carried out with respect to its Markov blanket, delimiting the sub-network of nodes that need to be considered. The Markov blanket consists of a node's parents (its direct ancestors) plus its children (its direct descendants) and the parents of its children. When the values of the variables within the Markov blanket are known, all other variables can be ignored: conditional on the nodes in its Markov blanket, any node is independent of all other nodes, regardless of the size of the network (Korb & Nicholson, 2011).

The second rule relating to d-separation, faithfulness, comes in two forms, strong and weak. The strong form states that if Z does not separate A and B, then A and B will be associated given Z. However, this rule may not hold, if, for example, there are two paths from A to B, representing contradictory causal relationships. If drinking alcohol shortens lifespan by its carcinogenic effects, and lengthens it by reducing stress, the expected relationship between A and B (alcohol and lifespan) may be unseen. Typically, the faithfulness rule is therefore expressed in its weak form, stating that given d-separation of A and B by Z, A and B *may* be associated given Z - a less satisfying but more reliable rule. In a simple common-effect network (1c in figure 1), the two causes are independent by definition (since if they have a shared cause which is not included in the CBN, the CBN is incomplete). Knowing the state of the effect means that each of the two causes is now informative about the state of the other cause.

In the next section, two important phenomena, discounting and augmenting, which derive from d-separation, and conditional d-separation, will be described more fully. They are central to the present research. When reasoners do not reason about them normatively, we can say that the Markov Condition, or the Markov Assumption, has been violated. There is however some inconsistency in terminology about such violations. As described above, the Markov Assumption applies to open paths d-separated by conditioning on an intervening variable. Using this terminology precisely, or restrictively, Derringer and Rottman (2018a), for example, discuss MA violations with respect to reasoning on common cause and chain networks (1a and 1b in figure 1), and they refer to non-normative reasoning with respect to common-effect networks (1c in figure 1) as failures of explaining away.

On the other hand, some authors do not consistently identify Markov Assumption violations so restrictively. Rottman and Hastie (2014), for example, review CBN-based reasoning research, and describe reasoning with both common-cause and common-effect networks in their discussion of the Markov assumption, and group failures to adhere to either compatibility or faithfulness as Markov Assumption violations. This laxer description is used here, and when discounting and augmenting are introduced, failures of both will be considered as possible Markov Assumption violations.

It is these requirements for accurate causal reasoning on open and closed paths, conditioning or not on an intercepting variable which are central to the current research, and which appear in everyday reasoning as the phenomena of discounting and augmenting. In a CBN framework, non-normative causal reasoning about such relationships is seen as violations of the MA. Such a violation is sometimes defined narrowly with respect to reasoning on structures lacking colliders. This is the case, for example, in Derringer and Rottman (2018a), where reasoners who reason non-normatively about a common-effect network are distinguished as failing to ‘explain away’, as described in the next section on discounting and

augmenting. For other authors, e.g. Rottman and Hastie (2014; 2016), failures of reasoning on all three types of network shown in figure 1 are Markov Assumption violations, and this broader terminology will be followed here.

5.5 Causal strengths

As well as parameterising a CBN via probability tables for each node, an alternative way to conceptualise the relationships between the variables is to see them as deriving directly from causal associations, with each edge having a causal strength giving the probability of the effect in the presence of the cause, and the absence of all other associations (Causal power theory: Cheng, 1997; Cheng and Lu, 2017). When considering CBN formalism as modelling human cognition, this conceptualisation corresponds to the position that people tend to reason about causes and effects, from which they derive statistical associations.

The formula in Cheng (1997) for causal power is

$$W = \frac{\Pr(q|p) - \Pr(q|\neg p)}{1 - \Pr(q|\neg p)}.$$

Causal power is "an unobservable enduring capacity to influence the occurrence" of the related effect. *All* the causes of an effect are partitioned into the cause in question (along with others on the same network path to the effect as the cause in question, and an amalgam of all other possible causes. If the cause in question is generative, then the causal power of a cause, c, is "the probability that e occurs due to c occurring" (Cheng & Lu, 2017, p. 7), where ‘e’ is the effect - this probability is unobservable. Causal power theory allows for preventive causal powers as well as generative, although in the present research all causes presented in experiments are assumed to be generative. In causal power theory the composite alternative cause is constrained to be generative.

Four assumptions are made (Cheng, 1997; Cheng & Lu, 2017, p. 7):

1 c and a influence e independently

2 a could produce e but not prevent it

3 the power of a cause is independent of the frequency of occurrence of the cause

4 e does not occur unless it is caused

A choice of an integration rule must be made to infer the probability of an effect in light of a cause and its alternatives. For a generative cause, the most commonly chosen rule is a 'noisy-OR' function.

Pr(effect = 1 | cause) = $1 - (1 - Wa)(1 - Wp)(1 - Wp)^{ind(p)}$

Here, the alternative causes are always present, as shown by the lack of a power on the (1 - Wa)term.

If, on the other hand, we wish to make inferences about a two-cause collider node in a CBN without including alternative causes, the relevant formula will be:

Pr(effect | cause 1, cause2) = $1 - (1 - W_{cause1}^{ind(cause1)})(1 - W_{cause2}^{ind(cause2)})$

In this formula, we are multiplying the possibility that the first cause fails to produce the effect (1 – the probability that it does) by the probability that the second cause fails to produce the effect, raising each term to the power of zero (i.e. = 1) if the cause is not operative. This product is the probability that both causes fail to cause the effect, and 1 minus that value is the probability of the effect.

In a case where an effect requires both causes, for example, plant growth requiring both light and water, a noisy-AND integration rule can be used:

$$Pr(effect = 1|cause1, cause2) \; = \; 1 - \; (1 - \; W_a)(1 - \; W_{cause1cause2})^{ind(cause1)ind(cause2)}$$

Chapter 6 Discounting and augmenting

6.1 Discounting and augmenting as intuitive ways to reason

The experiments reported here use materials intended to induce two phenomena in causal reasoning, discounting and augmenting. These can be described formally, most conveniently using CBN formalism. Discounting and augmenting are not obscure or difficult to understand, and when reasoning about everyday reasoning, we find these phenomena to be intuitive and easy to grasp.

When there are two independent causes of a single effect, one cause tells us something about the truth of the other when we know that the effect is true. Knowing one cause is true makes the other less likely: this is discounting. In everyday reasoning the rule is applied more effortlessly and ubiquitously than it is consciously understood, or applied. Thus, for example, the director Alfred Hitchcock could confidently make its use central to the plot of a number of his films. In Murder! (1930), Diana, an actress is accused of, tried, and found guilty of the murder of another woman, and sentenced to be hanged. When a fellow actor kills himself, leaving a note admitting the murder, Diana is freed (just in time). In Young and Innocent (1937), Robert flees from the police for a murder he didn't commit, until the real murderer confesses, and the police lose interest in Robert. To understand the plot of The Lodger (1927), The Wrong Man (1956), and Frenzy (1972), a viewer must recognise that the characters are using the same rule of reasoning. (https://en.wikipedia.org/wiki/Alfred_Hitchcock_filmography). These scenarios also point to an assumption noted later for common-effect networks, that the causes are independent. If the killer's suicide note in Murder! referred to 'my partner in crime, Diana', releasing the actress from prison would not have seemed so plausible as a plot development.

In reasoning research, discounting has also been known as 'explaining away' (e.g. Rehder & Waldmann, 2017), while when statistical inferences are biased by overlooking this effect, it

has been called 'endogenous selection' or ‘selection bias’ (Elwert & Winship, 2014). We can examine it without invoking CBNS by considering a pair of conditional sentences with a common consequent which is also a causal effect, such as:

If it rains, then the grass is wet

If the sprinklers are on, then the grass is wet

For the this pair of conditionals, when the truth of the effect is not known, the probability of one cause should not be affected by learning the other is true: for the pair above, rain and sprinklers on are independent in this case. However, if the effect is known to be true (the grass is wet), then one cause becomes less probable after learning the other is true: rain should be discounted when the grass is wet if we learn that the sprinklers are on. It has become a cause that is not needed to explain the effect.

The second phenomenon of importance for the current research is 'augmenting'. If a couple are in need of help in a foreign country, it would seem strange if one said to each other 'Let's ask that American - since he's reading an English newspaper, he'll probably understand us'. If the person in question is an American, why bother mentioning the newspaper as evidence that he is likely to understand English? Conversely, referring to a stranger whose nationality is unknown, the newspaper augments the probability that he understands English, and the same remark seems quite natural. Here, there are two effects, reading English and speaking English, which are both likely to be caused by 'being American'. Here the one effect makes the other more likely, but only if we don't know the status of the cause.

We expect augmenting if we have a pair of causal conditionals with a shared cause, the consequent, encoding a diagnostic argument:

For example, with this pair of diagnostic conditionals:

If it is warm outside, then it is sunny

If there are shadows, then it is sunny

For these conditionals, when the cause (the consequent) is known, the probability of one antecedent should not be affected by learning the other is true: for the pair above, warm and shadows are independent in this case, since either is already fully explained by the sun. If the cause is not known to be true, then one antecedent becomes more probable if the other is true: shadows become more likely if we learn that it is warm. Without knowing if the cause is true, either antecedent is a sign of the other.

To recapitulate, discounting is expected when there are two possible causes of an effect (e.g. two criminals for one crime) and the effect is known to be true, while augmenting is expected when there are two effects of a single cause (e.g. reading English and speaking English are effects of growing up in an English-speaking country) and the status of the cause is unknown. Making judgements from one cause to the other cause when we know the truth of the common effect means we are conditioning on the effect. Conversely, for augmenting, we judge the marginal probability of one effect from the presence of the other, i.e., unconditionally.

6.2 Formal accounts of discounting and augmenting

Causal reasoning about situations where discounting or augmenting is appropriate can be qualitatively or quantitatively normative. I will discuss what is normative with respect to the simplest appropriate networks, of three nodes, with a common effect or a common cause.

Discounting

In a common effect network (Fig 1A), with effect E, and causes C1 and C2, the simplest statement of discounting is the inequality

1] $Pr(C2 = 1 \mid E = 1, C1 = 1) < Pr(C2 = 1 \mid E = ?, C1 = 1)$.

In terms of the pair of common effect conditionals above, the probability that the sprinklers are on (*Pr*(C2)) given that the grass is wet (*Pr*(E) = 1) when it is also raining (*Pr*(C) = 1) is less than the probability that the sprinklers (Pr(C2)) are on given that the grass is wet (*Pr*(E) = 1) . Knowledge about the effect d-separates the two causes of rain and sprinklers, and makes the state of one no longer informative about the state of the other. If discounting does not occur, and the inequality does not hold, the Markov Assumption is said to be violated, since conditioning on the effect has not d-separated the two causes.

Augmenting

In a common cause network (Fig 1B), with cause C, and effects E1 and E2, augmenting is shown in the equality

2] $Pr(E2 = 1 \mid C = 1, E1 = 1) < Pr(E2 = 1 \mid C = ?, E1 = 1)$.

Using the common cause conditional pair above, the probability that it is warm (*Pr(E2*) given that there are shadows (*Pr*(E1)) is higher than the probability that it is warm (Pr(*E2*)) when it is sunny (*Pr*(C) = 1) and there are shadows (*Pr*(E1) = 1). Not knowing the state of the cause d-connects the two effects of warmth and shadows, and thus one effect is informative about the other. When the cause is known to be true, it screens off each effect from the other (just as knowing the state of the intermediate cause in a chain screens off the first cause and the effect from each other). If augmenting does not occur, the Markov Assumption can be said to have been violated, although some research reserves that description for an absence of discounting in common effect scenarios, and describes non-normativity in this case as a failure of screening off (e.g., Derringer & Rottman, 2018a).

Equations 1 and 2 are the same, except for the swapping of effects and causes, denoting the opposite causal directions of the common-effect and common-cause networks. If the

equations were written using the logical terms of antecedents and consequents, they would be identical. The distinct names, discounting and augmenting, denoting opposite valences, are due to a different choice of baseline for each equation. In each case the baseline is the case where the truth of one antecedent is irrelevant to the truth of the other, that is to say, when the status of the effect is unknown for common-effect networks, and when the status of the cause is known for common-cause networks. In CBN terminology, the baseline is when the three node network is d-connected, and discounting and augmenting refer to the difference between the baseline networks and their d-separated versions. For discounting, *(Pr(C2 = 1 | E = 1, C1 = 1) - Pr(C2 = 1 | E = ?, C1 = 1)) < 0*, and for augmenting, *(Pr(E2 = 1 | C =?, E1 = 1) - Pr(E2 = 1 | C = 1, E1 = 1)) > 0*.

Further, in equations 1 and 2, the truth of the consequent is denoted as unknown on the right side of each inequality. Assuming the consequent is not an impossibility, or a certainty, reasoners should assume that its probability is between 1 and 0. For each case, a three-part inequality can show what is expected in the case where the consequent is known to be false.

1a] $Pr(C2 = 1 \mid E = 1, C1 = 1) < Pr(C2 = 1 \mid E = ?, C1 = 1) < Pr(C2 = 1 \mid E = 0, C1 = 1)$

2a] $Pr(E2 = 1 \mid C = 1, E1 = 1) < Pr(E2 = 1 \mid C = ?, E1 = 1) < Pr(E2 = 1 \mid C = 0, E1 = 1)$

In real-life reasoning, this kind of expansion may not be applicable - for example, Hitchcock didn't make any films where the characters speculate on the likely murderer of someone who is not dead.

6.3 Discounting and augmenting and MMT

The following description of the MMT account of pairs of conditionals, for a common effect, and a for a common cause, and the associated discounting and augmenting is based upon that in Ali, Chater, and Oaksford (2011) and Hall, Ali, Chater, and Oaksford (2016). The faithfulness of such a description, based upon the principle of equal probabilities for each model represented was disputed by Johnson-Laird (2013), but this description holds to that principle for the reasons given in Hall et al. (2016).

For a pair of common effect conditionals, e.g.,

If she throws the vase, the vase breaks

If a tennis ball hits the vase, the vase breaks

discounting is normative (for CBN based accounts). That is to say, *Pr(throw|hit, break)* is less than *Pr(throw|break)*, while if the effect is unknown, *Pr(throw|hit)* is the same as *Pr(throw)* (i.e. the two causes are independent of each other).

According to MMT, such a pair of conditionals is represented as If cause 1 OR cause 1, then effect. Following the principle of truth, the three (out of eight) combinations of the two causes and one effect in which one or both causes are true, and the effect is false are not represented.

Thus, five models are constructed for the fully explicit model:

throw	hit	break
throw	¬hit	break
¬throw	hit	break
¬throw	¬hit	break
~~throw~~	~~hit~~	~~¬break~~

~~throw ¬hit ¬break~~

~~¬throw hit ¬break~~

¬throw ¬hit ¬break

For MMT, the equal probability of one cause given or not given the other cause does not hold. This is shown by comparing the proportion of models for each case. *Pr(throw)* = 2/5 (two models among the five). *Pr(throw|hit)* = 1/2 (one model among the two for which hit is also true). Conversely, where a CBN based theory predicts discounting when the effect is true, none is seen according to MMT. In this case, *Pr(throw|hit)* is 1/2, as we have just seen, and *Pr(throw|hit, break)* is also 1/2 (one model where throw is true among two where hit and break are true).

By contrast, analogous calculations for the common cause case, e.g., two conditionals such as

If there are shadows, it is sunny

If it is warm, it is sunny

show augmenting in the same case (cause true) as does a CBN-based account. Thus, the common effect case distinguishes between MMT and a causal account for pairs of conditionals.

Chapter 7 Markov Assumption violations in previous research

7.1 Overview

Representing causal relationships using CBN formalism has found practical use by making inferences about the causality computationally tractable (Korb & Nicholson, 2010; Neapolitan, 2004). In human reasoning research, a CBN can provide an explicit normative standard against which to assess causal reasoning. When reasoning is not normative, the CBN approach can provide an agreed description for deviations. Most commonly, an inference might be said to violate the MA, as described above.

Before examining some empirical results, it is worth noting some of the experimental practices they involve. Rottman and Hastie (2014) noted that two types of information are included in a CBN: the structure, i.e. the nodes and directed links, and the parameterisation, i.e. the marginal and conditional probabilities associated with the nodes (or the causal strengths associated with the links, and the associated integration rules). Rottman and Hastie found experimental participants were more often told verbally of causal structures than they had the structures demonstrated to them. Given only probabilistic information, reasoners usually failed to infer the associated causal structure (Mayrhofer, Goodman, Waldmann, & Tenenbaum, 2008), but were able to do so given extra information, such as the time order of the variables (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). As for the parameters, where they were not simply those of participants' prior knowledge, they were either described verbally or numerically, or presented via correlations which participants observed. Rehder and Waldmann (2017, p. 249), suggest that these differences are important enough to constitute a 'description-experience gap'. There are at least two ways in which experiments have typically used a subset of the scenarios which CBNs can represent. Firstly, the variables concerned have typically been binary. For example, people have been presented as pro-Castro or not (Morris & Larrick, 1995),

and ozone levels as high ozone or low ozone (Rehder, 2014), although CBNs allow for continuous variation in variable state. Secondly, while CBNs can represent both generative and inhibitory causes, almost all studies have been restricted to generative causes (Rottman & Hastie, 2014).

The empirical research described below is grouped, firstly, into results that compared human reasoners' inferences with the normative predictions of CBNs, secondly, results relating to evidence as to what causal relationships reasoners were actually making inferences about, along with particular conceptual nuances falling outside basic CBN representations, and, thirdly, results that throw light on possible heuristic strategies that reasoners might apply to causal scenarios. Two reviews of the empirical evidence are given by Rottman and Hastie (2014) and Hagmayer (2016).

7.2 Do human reasoners respect the Markov Assumption?

The success people show in navigating and interacting with a complex world is an indicator that they are skilled in causal reasoning. Numerous studies have found abilities in reasoners such as understanding intervention versus observation, taking account of base rate, and distinguishing causal relationships according to their strength (Hagmayer, 2016).

Studies with scenarios which imply discounting (i.e., for which the CBN includes a collider) have most often found the effect to be absent, or weaker than expected (e.g., Davis & Rehder, 2017; Morris & Larrick, 1995; Rehder & Waldmann, 2017; Rottman & Hastie, 2016). For example, Morris and Larrick, (1995), told participants of students who had been instructed either to write an essay praising Fidel Castro, or instructed to write an essay either in praise of or attacking the Cuban leader. Participants who were told that a positive essay they had read had in fact been written by a student instructed to praise Castro should have discounted the

essay as evidence that its author was pro-Castro. The discounting effect was weaker than it should have been.

The phenomenon of augmenting, as described above for common-cause networks, is an example of 'screening off', found for three-node networks lacking a collider, that is to say, common-cause and chain structures. Results suggesting Markov Assumption violations have been found for these types of network too, for example, Walsh and Sloman (2004), using common-cause scenarios, and Chaigneau, Barsalou, and Sloman (2004) for chain scenarios. Rehder (2014) presented common-cause relationships where the levels of the variables involved (low or high) were presented in different combinations, with the intention of reducing the effects of prior knowledge of the topics among participants. Although a majority of judgements were normative, over 30% were not.

7.3 Do reasoners and experimenters consider different causal networks for specific scenarios?

Any claim of a Markov Assumption violation has to be made with respect to a particular causal structure. If a reasoner considers a causal network that is different from that assumed by an experimenter, the reasoner may not have violated the MA, but be making normative inferences. This has been generally recognised by researchers. The experimenter sees an apparent violation due to ‘an incomplete causal graph’ (Hagmayer, 2016). Obviously, "it is very hard to know which causal structures people are actually using" (Park & Sloman, 2013). “It is fairly plausible to assume that with real-world scenarios experimental subjects do not always stick to the instructions provided in the cover stories but tend to augment the instructed model with additional hidden variables coding their background knowledge” (Mayrhofer & Waldmann, 2014).

Walsh and Sloman (2004; 2007) addressed this question directly. They looked at common cause networks, and found that their participants, when asked, described causal networks with which the participants' inferences were compatible, but differing from those presented by the experimenters.

Walsh and Sloman asked participants to reason about real-world common cause relationships. For example, in one condition, worrying (C) caused poor concentration (E1) and insomnia (E2). In another, jogging (C) caused improved fitness (E1) and weight loss (E2). Participants were asked to give the probability of E2 given C, and of E1 given C and not-E2. In the corresponding CBN, conditioning on the cause d-separates E1 and E2, making the extra information in the second question irrelevant. Participants, nonetheless, judged that knowing, for example, that someone did not have insomnia reduced the probability of poor concentration for a worrier.

How participants viewed the relationships in these scenarios was examined in a further experiment. The responses revealed that participants typically added, from their world knowledge, additional shared causes (generative or inhibitory) of the two effects. For example, jogging increased appetite, inhibiting both weight loss and improved fitness, or a lack of insomnia was due to relaxation exercises, which also improved concentration.

Mayrhofer, et al. (2008) also used common cause scenarios to investigate apparent Markov Assumption violations in reasoning, and influenced participants judgements by manipulations intended to make effect correlation more likely.

In a three-node common cause network, if we expect the cause to produce both effects, we may see the cause being true without a particular effect being true as a failure of the causal relationship. Although the effects are independent of each other in such a network, conditional on the cause, extending such a network to account for such failures might take place in two

ways: there might be two separate, and independent, disabling causes, or a single disabling cause, acting on both of the cause -> effect relationships simultaneously. Once the simple network is extended in such a way, there might be no Markov Assumption violation.

Mayrhofer et al. (2008) presented participants with scenarios involving mind-reading aliens of green or yellow colour. Different degrees of Markov Assumption violations suggest that the experimental manipulation induced extending the networks to give common causes, and thus correlated errors of thought transmission, within, but not across, the alien colour divide. Markov Assumption violations were observed more in scenarios where the aliens were of the same colour, suggesting that participants were more likely to consider correlated errors, from a cause beyond the causal structure presented, for mind-reading results across aliens of the same colour.

Mayrhofer, Hagmayer, and Waldmann (2010) also attempted to induce participants to extend common effect networks in these ways. They presented scenarios with mind-reading (extra-terrestrial) aliens, where the thoughts of one alien (the cause; C; called “Gonz”) became present in the minds of three others (the effects; E1, E2, E3; called “Murks, Brxxx, and Zoohng”). The manipulation which aimed to cause participants to see failures to transfer thoughts as independent or not was the description of the process as 'reading' or 'sending'. In the reading condition Markov Assumption violations were smaller than in the sending condition. Plausibly, in the reading condition, the three thought-receiving aliens can be seen as active agents, using three separate processes to access Gonz's thoughts. If this is the case, it seems plausible to see the effects as independent of each other, and failures as due to the individual characteristics of the receivers. Thus, this extended network, with an separate disabling cause for each mind-reader does not easily lead to violations of the conditional independence from the MA. By contrast, if the story presented assigns the thought-transfer agency to the actions of Gonz, the sender, failures are more likely to be located in a single

shared process, and thus the states of the three receivers remain correlated even when Gonz's mental state is known. These verbal manipulations were inspired by a 'dispositional' account of causality (Wolff, 2007). Typically, a cause is represented linguistically as having agency (e.g., 'the robber showed the bank teller his gun') and an effect is denoted as a 'patient', acted upon by the agent, but this can sometimes be reversed (e.g. 'the bank teller looked' versus 'the bank teller saw'). These are aspects of causal cognition that are do not have a place in the basic CBN formalism.

These results show the importance of both background knowledge and the words with which a scenario is presented on reasoning, and suggest that experiments should either address these factors explicitly, or analyse the data using statistical methods intended to reduce the effect of the particular scenarios used, and how they are presented.

7.4 Do reasoners generally fail to adhere to the MA?

Another approach has been to look for general characteristics of causal reasoning that may affect all judgements, regardless of the content. These might extend to abstract, content-free, reasoning tasks, for which Markov Assumption violations have also been found.

Rehder and Burnett (2005), and Rehder (2014) proposed a more general explanation for Markov Assumption violations in the causal reasoning of their participants in the form of a general propensity to assume an 'underlying mechanism' not grounded in particular scenarios or their presentation. Rehder and Burnett (2005) presented participants with category-based tasks. Imaginary constructs were to be categorised according to their features, and these features formed part of networks showing causal relationships among them. The categories ranged from (imaginary) biological species to abstract materials with letters as features.

Although Rehder and Burnett's networks had four nodes, they fell into the three distinct minimal three-node groups: common effect, chain, and common cause.

Participants' judgements were not in line with the MA. For example, for a common cause network (C -> E1, C -> E2, C -> E3), present effects were seen as making a different effect more likely when the state of the common cause was known. This is a Markov Assumption violation: in this network without a collider, the d-connected path becomes a d-separated path when an intervening variable (C) is conditioned on, and any E should become independent of the other Es.

For a chain (C1 -> C2 -> C3 -> E), the value of the effect (E) variable was still affected by the value of the first two (C1 and C2) nodes in the chain when the value of the penultimate node (C3) was given. This is a Markov Assumption violation: in this network without a collider, the d-connected path becomes a d-separated path when an intervening variable (C3) is conditioned on, and E should become independent of C1 and C2. Similar results obtained for reasoning in the diagnostic direction, i.e., inferring C1, given C2.

For a common effect network (C1 -> E, C2 -> E, C3 -> E), the value of one cause (C1, C2, or C3) was dependent on the values of the other causes when the status of the effect was unknown. This too is a Markov Assumption violation: a collider node intervening on a path d-separates the nodes it lies between, which are thus marginally independent, though the path is opened if collider variable is conditioned on.

From a CBN viewpoint, the explanation (or justification) for these Markov Assumption violations which was proposed by Rehder and Burnett (2005) is simple: they suggested that all four nodes corresponding to the category features they gave in the experimental scenarios were represented by participants as downstream from a further *shared* cause (a fifth node), which Rehder and Burnett called an 'underlying mechanism'. The question is whether it would be reasonable for participants to extend the causal relationships in this way.

Here the range of domains represented by Rehder and Burnett's categories is important. While a 'biological essence' underlying the features of (imaginary) species may be intuitively plausible, it is hard to see how participants could believe they have experience with categories ("Daxes") whose features are letters (A, B, C...) could be justified in inserting a shared mechanism for those letters. Thus, Rehder and Burnett proposed a domain-general causal reasoning propensity, by which people assume hidden, shared, causes behind natural phenomena, and apply this heuristic even to causal relationships between abstract entities.

Rehder (2014) used (invented) non-categorical relationships, and counterbalanced the direction of the causal relationships in an attempt to stop participants applying pre-existing world knowledge. Similar results were again found.

In a three-node network representing two causes of a single effect (and also for similar networks with three or more causes), the network is asserting that the two causes are independent of each other. If we choose to represent a common-effect scenario by such a CBN, we are assuming that the causes we represent are independent. Such a network structure must be justified by the structure of the world. Consider, for example, two causes of wet grass: rain, and functioning sprinklers. Is this really a causal scenario in which we can be confident that the two causes are independent? That is to say, does whether it is raining or not raining really not give us a reason making some assumption about the probability of sprinkler use? And while our understanding of the causal structure of the world surely makes us confident that the state of sprinklers is not a cause of the state of rain, can we really say that seeing sprinklers on all over a neighbourhood as we pass through does not make some (diagnostic) inference about the weather forecast, and from that infer something about the likelihood of rain?

Von Sydow, Hagmayer, Meder, and Waldmann (2010) showed that, regardless of any propensity of reasoners to assume correlated causes, experimental participants were able to

learn that causes are independent. Participants were told a two-cause, one-effect, causal relationship, and then presented with repeated trials allowing them to learn the associations between the three variables. In these trials, the two causes were not associated, and participants judgements after the learning trials showed that they had successfully learned this independence.

Rottman and Hastie (2014, p. 116) (2014; p. 116) reviewed a series of studies on when, and whether, reasoners inferred that one cause (C1) of an effect was evidence for the other (C2), and found results which were contradictory, not merely between studies, but in one case within a single study (Hagmayer and Waldmann, 2007). Rottman and Hastie conclude that "people's beliefs about the relationship between C1 and C2 are inconsistent and in one instance go against the normative framework".

A number of researchers have proposed that causal judgements which are non-normative may be due to reasoners privileging in some way two particular network states: the state in which all variables are true, and that in which all nodes are false. For the three-node networks which are the simplest for which the Markov Assumption is important, these states are a quarter of the possible network states. For larger network the proportion becomes smaller, and thus the bias larger, with the size of the network. If a reasoner infers the state of a variable given information about some other nodes, and if a majority of those nodes is in one state, the reasoner may infer that the node in question matches the most common state among those for which information is available. If a reasoner does so, the Markov Assumption will be ignored, and is likely to be violated.

This proposal has been put forward under various names; for example, "associative bias" (Rehder, 2014), a "monotonicity assumption" (Rottman & Hastie, 2016), the "rich-get-richer principle" (Rehder & Waldmann, 2017), and "cue consistency" (Derringer & Rottman, 2018a).

A similar bias lies behind the discrete sampling account of CBN-like reasoning put forward by Davis and Rehder (2017; 2020).

Although a major advantage of using CBNs to represent causal information for artificial intelligence is that the explicit encoding of direct causal independence (via absent links) makes the calculations tractable, in practice many applications nevertheless require heuristic approximation due to hardware / time limitations. MCMC methods are commonly used. Davis and Rehder (2020) propose an adaptation of an particular MCMC algorithm (Metropolis-Hastings) as a algorithmic-level (Marr, 1982) account of what reasoners might be doing for causal inferences. Two restrictions make the computations required simpler, and affect the results in ways that resemble those found empirically. Each new proposed state to be considered differs from the current state in only one variable, and the starting point for the sampler can only be one of the two states where all variables have the same (binary) value. The effect of these restrictions is greater the fewer times the network is sampled. While every use of an MCMC algorithm uses a finite number of samples, Davis and Rehder (2020) suggest cognitive limits imply small sample numbers, increasing biases.

Three studies which have not been mentioned in this overview of the Markov Assumption in empirical reasoning research will be discussed together in the next section, as they are the direct background to the present research.

Chapter 8 The response mode discrepancy in previous research

8.1 Overview

The experiments to be reported below were carried out to learn more about a pattern of judgements for causal conditional scenarios reported in Ali et al. (2011). Pairs of conditionals elicited differing judgements across two ways of producing and giving responses, although the conditionals, and thus the logical and causal structure they asserted, were the same. Neither a logical nor a causal account of reasoning seems able to explain this difference.

Participants were told of common-effect and common-cause scenarios. They were given information about one of a pair of causes, or a pair of effects, and asked about the probability of the other cause, or effect. Also asserting the truth of the shared effect, or not asserting the truth of the shared cause, made discounting or augmenting normative.

The novel innovation in Ali et al. (2011) was the use of two parallel response modes. In one mode, participants were asked to give a single judgement as to a qualitative change in probability. In the other, they were asked to give two separate assessments of the probability of one cause / effect before and after being told of the other. The sign of the difference of those two values should have corresponded to the answer given in the other response mode. Ali et al. found that the results of the two modes did not match. These two response modes are referred to as the change mode, and the delta mode, respectively, and are explained in more detail below.

Replicating the results of that research, and especially the response mode difference, and examining possible explanations for it, is the goal of the present research.

Ali et al. (2011) forms part of a brief series of related earlier studies, of which is it the only study to present both response modes. Ali, Schlottman, Shaw, Chater, and Oaksford (2010) used the delta mode only. Hall et al. (2016) used only the change mode. The main supervisor of this thesis, Mike Oaksford, is an author on all three of these studies. The contents of the first

two of these papers will next be briefly summarised, along with a paper by a different team which also uses (though naming them differently) the two response modes: Tešić, Liefgreen, and Lagnado (2020). A fuller description of the research reported in Hall et al. (2016) will be given separately below, in chapter 9.

8.2 Ali, Schlottman, Shaw, Chater, and Oaksford (2010)

In the 2010 study, Ali et al. used child participants. Due to the young ages (6 to 8 years old) of the participants, the causal scenarios were presented not in writing, but rather demonstrated by means of specially-prepared wooden boxes with pictures, lights, and buttons. The experimenter also told the children the conditional relationships that they were being shown. For example, the two relationships of the common-effect scenario:

If the sun is shining, then the flower sparkles

If there is water, then the flower sparkles

were demonstrated with buttons for the antecedents, and a sparkling light for the consequent. The light was covered with a scarf for the 'no consequent' condition.

This experiment used what will be called here the 'delta response mode'. Participants gave an initial rating of one antecedent's likelihood, and then, after being told, or shown, that the other antecedent was present, participants gave a second rating. The second rating minus the first rating was the dependent variable, deltaR, negative when participants discounted and positive when they augmented. The children gave their ratings by pointing to the corresponding position on a shaded linear scale.

With 2 causal directions (common effect, 'CE', where CE stands for 'cause-effect', i.e. predictive, and common cause, 'EC', where EC stands for effect to cause, i.e. diagnostic) each presented either with the consequent known or unknown, the experiment had four conditions.

These four conditions, which are fundamental to the experiments described here, will be denoted by appending ‘C’ (consequent present) or ‘NC’ (consequent absent) to the abbreviations for the two causal directions. Thus, the four conditions in these three papers are CE-C, CE-NC, EC-C, and EC-NC. When reasoning causally, discounting is normative for one condition, CE-C, and augmentation also for one condition, EC-NC.

Ali et al., (2010) found that their child participants reasoned normatively in 3 of these 4 conditions, that is to say, for CE-C, where they discounted, and for EC-NC, where they augmented, and for EC-C, where they did neither (i.e., deltaR, or R2 – R1, was not significantly different from zero). For the fourth condition, CE-NC, there was a failure to reason normatively, since the participants discounted (or, more precisely, showed discounting-like behaviour) in this condition. That is to say, the children felt that one cause made the other less likely even when they did not know the status of the effect. Since, in a CBN, the path between the two causes is blocked unconditionally, when the consequent is not known, the causes are independent. This then, was a Markov Assumption violation, in CBN terminology.

8.3 Ali, Chater, and Oaksford (2011)

Ali et al., (2011), reported two experiments, with a larger number of causal scenarios. Their second experiment included a second response mode and revealed the discrepancy which is important to the present research. In experiment 2, participants were shown the scenario presented with or without the assertion of the conditional. In the description, the participant was told 'you wonder whether' the first antecedent (cause or effect) is true, and then, after the other antecedent is asserted, the participant was asked if they 'now' think the first antecedent is less, equally, or more likely - a forced choice among three responses. The information about

the first antecedent was additionally, in the consequent (C), condition, accompanied by a statement that the shared consequent was true.

This ‘change response mode’ asked participants to directly give their qualitative judgement of whether they thought discounting or augmenting (or, more precisely, discounting-like behaviour or augmenting-like behaviour) was appropriate, or neither was.

The dependent variable for this response mode was referred to as the ‘change rating’, abbreviated to CR. The dependent variable for experiment 1, and the earlier experiment in Ali et al. (2010), was called the ‘delta rating’, abbreviated to DR.

The conditional sentences presented to participants in Ali et al., (2011) can be interpreted both as logical and as conditional statements. The terms can be symbolised either as P1, P2, and Q (on the basis of the logical form of the conditionals, where P is an antecedent, and Q is the consequent), or as C1, C2 and E or E1, E2 and C (on the basis of the causal form of the conditionals, where C is a cause, and E is the effect). Thus the (predictive, CE) conditionals are of the form

If P1 / C1, then Q / E

If P2 / C2, then Q / E

The (diagnostic / EC) conditionals are of the form

If P1 / E1, then Q / C

If P2 / E2, then Q / C

Three ways in which both these experiments differed from Ali et al., (2010), were: 1) the participants were adults, not children, 2) the conditionals were presented in writing to the participants, rather than being demonstrated and explained by the experimenter, and 3) C-NC

was a between-subjects manipulation, rather than within-subjects as for the earlier experiment, where participants first judged the scenarios in the NC condition, then in the C condition.

In experiment 1, using the delta rating response mode, Ali et al., (2011) found the common effect scenarios elicited results in line with those of Ali et al., (2010), for the CE-C condition, for which discounting was found, but for the CE-NC condition, there was no discounting-like behaviour. For the diagnostic scenarios, augmenting was again found for EC-NC. The EC-C condition produced discounting-like (non-normative) judgements. (Distinguishing the terms 'discounting' and 'discounting-like' behaviour can help clarify the result of these experiments.)

In Ali et al., (2011), the results for experiment 2, using the change response mode, were unexpected: augmenting for EC scenarios regardless of the assertion of the cause (C or NC), and discounting for CE scenarios regardless of the assertion of the effect (C or NC). That is to say, participants preferred to judge the first effect as 'more likely' after hearing that the second effect was true, and their judgements were not affected by hearing the cause was true for pairs of diagnostic conditionals. For EC scenarios, participants judged the first cause as ‘less likely’ after the hearing that the second cause was true, and their judgements were not affected by hearing the effect was true.

8.4 Tešić, Liefgreen, and Lagnado (2020)

The origin of the present research lies in two studies introduced above, Ali et al., (2010) and Ali et al., (2011), which first revealed the response mode discrepancy. Hall et al., (2016) reported two experiments carried out in response to those earlier studies, with the aim of replicating the change mode result, and examining a possible explanation for the discrepancy. This paper shared with Ali et al., (2010) and Ali et al., (2011), (and also Hall et al., (2016), to be discussed in the next chapter) a common author in Mike Oaksford, the main supervisor of this PhD thesis.

The two response modes also appear in a paper, Tešić et al., (2020), which does not form a link in the above chain of related research. I will next give a summary of the experiments in that paper, and the results. Later, in the general discussion section, I will consider the explanations proposed in Tešić et al., (2020) (research itself related to an earlier study, Liefgreen et al. (2018)) for non-normative conditional reasoning behaviour.

Liefgreen et al., (2018) used three-variable common-effect conditional scenarios, corresponding to those used in the present research, as well as scenarios with five variables, i.e. two effects, each with two causes, one a common cause for the two effects (thus the scenarios corresponded to 3 and 5 node CBNs). Each scenario was presented with prior probabilities of the causes (i.e., their baseline, unconditional probabilities) at 'medium' and 'low' values (0.5 and a combination of 0.2 and 0.1). The scenarios all gave tossed coins as causes of the lit / unlit status of lightbulbs as effects.

Liefgreen et al., (2018) found that many participants (in one of the four conditions, a majority), when asked for the probability of a cause given the effect, or the effect and the other cause, reported the baseline probability of the cause regardless of the status of the effect, and Liefgreen et al. suggested an explanation for this: that these participants were using a 'propensity' interpretation of probability.

Tešić et al., (2020) built upon Liefgreen et al., (2018) in three important ways. Firstly, this new study replicated the pattern reported in the earlier study, whereby some participants fail to take the status of an effect when reporting the probability of a cause. This is the participant behaviour that Tešić et al., (2020) and Liefgreen et al., (2018) call the 'propensity interpretation'. Secondly, Tešić et al., (2020) give a new explanation for non-normative judgements by a second group of participants, a judgement pattern which they call 'diagnostic split', whereby

reasoners assign probabilities to a pair of causes such that their sum is 1. In Tešić et al., (2020) the participants from the 2018 study are discussed as falling into two clusters, one of which showed this pattern of reasoning (despite this description Tešić et al., (2020) did not carry out a cluster analysis to reveal these groups). Thirdly, Tešić et al., (2020) introduced two new scenarios, one of which is specifically intended to reduce the use of what Tešić et al., (2020) and Liefgreen et al., (2018) have called the 'propensity interpretation' by participants.

Experiment 1 of Tešić et al., (2020) uses three scenarios. The first, carried over from Liefgreen et al., (2018), describes two rooms, each with an 'automated coin tossing mechanism', and a 'storage unit' containing a light bulb, connected by cables to the coin tossing mechanisms. The only human involved is the participant, 'you' in the text shown to participants. The coin tossing mechanisms are described as being able to be set to produce a specific 'chance' of producing a head (described as a percentage, with the chance of a tail given as the complementary percentage). The light bulb turns on, deterministically, unless two tails are produced.

The participant is described as moving between the rooms and the storage unit, making observations, and is asked to give probability judgements on the basis of those observations. There is one exception: before being asked for the $Pr(C1 = \text{head} \mid E, \neg C2 = \text{tail})$ the participant is asked to 'imagine a scenario'.

The second scenario describes two containers with small balls, in stated proportions of rubber and copper balls, such that a ball is randomly chosen from each container to be inserted in an electrical circuit, lighting up a light bulb unless both chosen balls are rubber.

The third scenario describes a dinner party, where three invitees are named: Michael, Tom, who have not yet arrived, and Helen, who wants to drink red wine. The participant, 'you', who is the host, messages the two yet-to-arrive guests individually asking each to bring red wine.

The probabilities that they will do so are stated. In this scenario, Michael bringing red wine and Tom bringing red wine are the causes. Helen drinking red wine is the effect.

Thus. an important difference between Tešić et al., (2020) and Liefgreen et al., (2018) on the one hand, and Ali et al., (2010), Ali et al., (2011), and Hall et al., (2016), and the present research on the other, is that in both Tešić et al., (2020) and Liefgreen et al., (2018), participants were told the relevant unconditional probabilities (and then asked the same probabilities, as a check). In all the other studies scenarios were presented via conditional sentences, but associated probabilities were not suggested to participants by the experimenters.

Tešić et al., (2020) asked their participants for the following probabilities (experiment 1) (e.g., for the coin scenario, true = a coin is heads / the light bulb is lit, for the causes and the effect, respectively):

1 *Pr*(C1 = true) as percentage

2 *Pr*(C2 = true) as percentage

3 If C1 = 'were' true, would *Pr*(C2 = true) increase / decrease / stay the same

4 If C2 = 'were' true, would *Pr*(C1 = true) increase / decrease / stay the same

(participants were asked to write an explanation for their reasoning behind the answers to questions 3 and 4)

5.1 Given E = true, would *Pr*(C1 = true) increase / decrease / stay the same compared to question 1

5.2 Given E = true, *Pr*(C1 = true) as percentage

(participants were asked to write an explanation for their reasoning behind the answers to question 5)

6.1 Given E = true, would *Pr*(C2 = true) increase / decrease / stay the same compared to question 2

6.2 Given E = true, *Pr*(C2 = true) as percentage

(participants were asked to write an explanation for their reasoning behind the answers to question 6)

7.1 Given C2, E = true, would *Pr*(C1 = true) increase / decrease / stay the same compared to question 5

7.2 Given C2, E = true, Pr(C1 = true) as percentage

(participants were asked to write an explanation for their reasoning behind the answers to question 7)

8.1 Imagine that E = true, but instead of C2 = true, C2 = false, would *Pr*(C1 = true) increase / decrease / stay the same compared to question 5

8.2 What do you now think is *Pr*(C1 = true) as percentage

(participants were asked to write an explanation for their reasoning behind the answers to question 8)

When classifying qualitive responses as normative, Tešić et al., (2020) accepted answers within a bracket of plus or minus 0.2.

Although the three scenarios, and the rationale behind them, are of considerable interest, Tešić et al., (2020) found that for discounting (which they named 'explaining away') in both modes, there was no main effect of scenario, nor an interaction effect with the level of prior probability given to the participants, and thus they did not include scenario in their analyses of

the responses relating to discounting, retaining only the three levels of unconditional probability of cause (low / medium / high prior).

Looking at the results corresponding to the CE results for Ali et al., (2010), Ali et al., (2011), Tešić et al., (2020) found insufficient discounting. For the change mode, ('qualitative' in Tešić et al.), correct response percentages for the three prior levels were 36%, 21.7%, and 19.9% (low / medium / high). For the delta mode ('quantitative'), correct response percentages for the three prior levels were 52.7%, 82.9%, and 57.6% (low / medium / high). Tešić et al., (2020) discuss this difference, suggesting that the comparatively high level of normative explaining away in the delta (quantitative) mode may be deceptive, due to participants giving the unconditional probabilities of the causes in their responses (the priors). From the viewpoint of Tešić et al., (2020), the change / quantitative responses are a useful sanity check on the probabilities given by participants.

Tešić et al., (2020) is interesting for the grouping of participants into two ‘clusters’, each associated with a possible explanation for its observed reasoning pattern. One group of participants is assumed to be carrying out a 'diagnostic split' reasoning heuristic, whereby these participants preferred to assign probabilities for the two causes given the effect, which summed to 1. The second cluster of participants tended, when asked for the probability of one cause given the effect, or one cause given both the effect and the other cause, to give the prior probability of the cause, that is to say the unconditional, background probability of the cause, as told to them during the presentation of the scenario. Tešić et al., (2020) suggest that these participants are showing a 'propensity interpretation' of the relevant probabilities, where 'propensity' refers to the interpretation of a probability value: not the limiting value of an infinite number of trials, and not a degree of belief, but an inherent property of a particular physical system or mechanism. This interesting hypothesis is new with Tešić et al., (2020), and has implications beyond the focus of the present study, which is on replicating Ali et al., (2010),

Ali et al., (2011)'s results on discounting and augmenting. I will return to look further at the 'propensity interpretation' in the general discussion.

8.5 What is to be replicated?

The primary goal of the experiments to be reported below is to see if the results of Ali et al. (2010, 2011) and Hall et al. (2016) replicate (a summary of these results is shown in table 1, and model graphs are shown in figure 2). For each causal direction, in all experiments, and both response modes, discounting was found for common effect conditional pairs (i.e., calling for predictive reasoning) when the consequent was present. Similarly, for both causal directions, every experiment and mode produced augmenting for common cause conditional pairs (requiring diagnostic reasoning) when the consequent was not stated. These are consistent, and also consistently normative, results.

Table 1

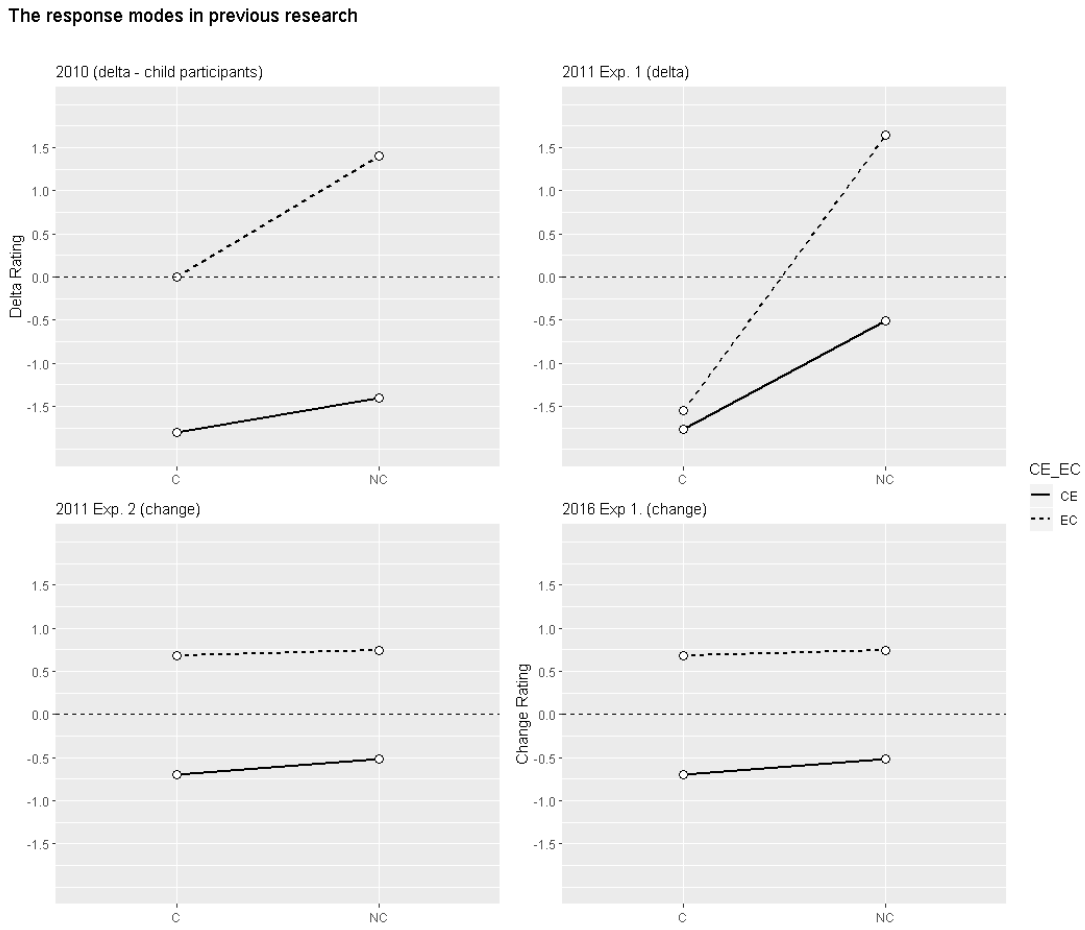
The Effects Found by Ali et al., (2010, 2011)

	Normative	2010 (delta)	2011 delta	2011 change	2016 (change)
CE-C	Discounting	Discounting	Discounting	Discounting	Discounting
CE-NC	Neither	Discounting	Neither	Discounting	Discounting
EC-C	Neither	Neither	Discounting	Augmenting	Augmenting
EC-NC	Augmenting	Augmenting	Augmenting	Augmenting	Augmenting

The effects found by Ali et al., (2010, 2011), described in this section, and Hall et al., (2016, experiment 1) described in the next chapter. Normative results for causal reasoning are shown in grey. For graphs of model values see figure 2.

Figure 2

Results of Previous Research Demonstrating the Response Mode Discrepancy



Results of previous research demonstrating the response mode discrepancy: see next chapter for a discussion of Hall et al., 2016

The response mode discrepancy was found by Ali et al. (2011). Their change mode results showed, in both causal directions, that the two conditions where asserting one consequent should not affect the probability of the other (CE-NC and EC-C) produced judgements for which the truth of one consequent did affect the probability of the other. The participants did not take into account the status of the antecedent of the conditionals in their reasoning. The experiments reported here will demonstrate if these results replicate.

However, a further discrepancy, this time between experiments, occurred in the earlier research for the delta mode. While the experiments for the change mode (Ali et al., 2011, experiment 1; Hall et al., 2016, experiment 1) both used adult participants, those for the delta mode (Ali et al., 2010; Ali et al., 2011) used children in one, and adults in the other, and it is reasonable to say that neither result has been replicated so far. The children demonstrated inappropriate discounting-like behaviour for the CE-NC condition, in line with both change mode results, but different from the normative adult judgements for CE-NC. Conversely, the children reasoned normatively in the EC-C condition, a different result from all change mode results in this condition, and also different from the adult judgements. For EC-C, in the delta mode, the adult participants in Ali et al., (2011) did not merely reason non-normatively for the consequent present condition, but did so by revealing discounting-like behaviour for diagnostic conditional pairs. If participants were representing this condition with a CBN, it appears they were reversing the causal direction, treating EC-C as CE-C, and thus discounting. The experiments reported in this research, using adults throughout, should reveal if these are cases of superior judgement by younger participants.

Chapter 9 A first replication and a first explanation – findings reported in Hall, Ali, Chater, and Oaksford (2016)

9.1 The rationale

In Ali et al., (2011), Experiment 1 (delta mode) is a first attempt at replicating, with a somewhat different method, and leading to somewhat different results, Ali et al., (2010). Experiment 2 of Ali et al., (2011) introduced the change mode. A first attempt to replicate the change mode results of Experiment 2 of Ali et al., (2011) was reported in Experiment 1 of Hall et al., (2016).

The results of that experiment, and the analysis examining if the results supported causal model theory, or rather supported mental model theory will be given here, based on the original publication. Also reported in that study, and reported again here, is Experiment 2 of Hall et al., (2016), which examined a 'shallow encoding' explanation, due to Mike Oaksford, for the change mode results, introducing new scenarios designed to test that explanation. A somewhat fuller report, including an introductory discussion of topics mostly covered in the introductory chapters above, is to be found in Hall et al., (2016).

Three broad background frameworks for understanding human reasoning with conditionals were described above: traditional logic, mental models, and the probabilistic / causal framework. For experiment 1 in Hall et al., (2016), different predictions of the results were made in line with mental models and causal models.

As described above (chapter 2, section 2), mental models theory predicts reasoning will differ according to the time and effort reasoners can allocate to a task. Given time, inclination, and mental resources, a reasoner can produce a 'full model' (FM), including a representation of each possibility compatible with the information available. Alternatively, a reasoner may merely produce an 'initial model', (IM), with an incomplete set of representations. To compare

mental models and causal model (CM) theories, Hall et al., (2016) made three predictions with respect to these three reasoning strategies about the results of their change mode experiment (see Figure 3):

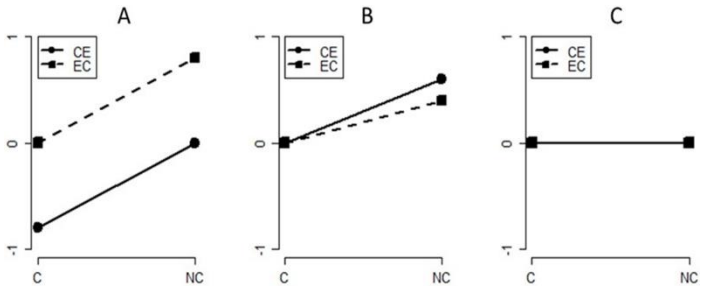
1) For CM, both causal direction (causal (CE) vs. diagnostic (EC)) and assertion of consequent (present (C) vs. absent (NC)) will give main effects. For mental models, in the FM version, assertion of consequent will give a main effect, while neither consequent or causal direction will lead to a main effect if reasoners produce only an IM.

2) For CM, for the CE-C condition (causal direction, consequent present) the mean change rating should be less than in the EC-C condition and below zero. This is discounting, which will not be produced by either mental model version.

3) For CM and FM, for the EC-NC condition (diagnostic direction, consequent absent) the mean change rating should be higher than in the CE-NC condition and above zero. This is augmenting, which will not be produced in the IM version of mental models.

Figure 3

Predictions for Causal Models, Full Mental Models, and Initial Mental Models



Predictions for causal models (A), full mental models (B), and initial mental models (C). The figure is taken from Hall et al., (2016). Average predictions for common-effect (CE) and common-cause (EC) conditional pairs, with (C) and without (NC) the consequent asserted. Positive values represent predicted responses of 'more likely, negative represent 'less likely' and 'equally likely' is given a value of zero.

9.2 Pre-test of materials

A pre-test experiment was carried out to check the suitability of the materials for each of the two main experiments. This pre-test was carried out by a separate group of participants (N = 18).

Participants were shown phrases corresponding to the three variables in pairs of common-cause or common-effect variables (not positioned consistently on the paper as either antecedent / consequent or cause / effect) and were asked to draw arrows for the causal connections they considered valid. They then rated the strength of each causal connections by writing a number from 1 to 4. The pre-test data was used to select materials from an initial pool of 20 pairs of causal / common-effect / CE conditionals, and 13 pairs of diagnostic / common-cause / EC conditionals for experiment 1, and a pool of ten causal, and nine diagnostic / common-cause / EC conditional pairs for experiment 2. Each pair of conditionals was embedded in a scenario, to render the inference participants were being asked to make more realistic and intuitive. Pairs of conditionals from the initial pool were excluded:

1. If the causal directions were not unidirectional.

2. For causal pairs if both causes failed to have a similar causal power, and for diagnostic pairs if the common cause failed to have a similar causal power for both effects.

9.2.1Pre-test participants

In the pre-test phase, 18 undergraduate students (3 male, 15 female) from University College London took part (mean age: 24.1 years, range: 18-53 years).

Pre-test procedure. In the pre-test, participants were asked to draw arrows from the statements that they thought were the causes to the statements that they thought were the effects. Next participants were asked to indicate how strong they thought the causal relationship was between a particular connection on a scale of 0-4; (0 for a very weak causal relation, 1 for weak, 2 for average strength, 3 for strong and 4 for very strong). One set of materials appeared per page of a booklet and the pages of each booklet were randomized.

9.2.2 Pre-test results and discussion

The pre-test data was coded so that 0 = no causal relation, i.e., no causal link inserted into the diagram. Thus, the causal rating scale was rescaled to a six-point scale from 0 to 5. There were two expected causal relations in each of the 33 pairs of conditionals tested. For CE these were p1 → q and p2 → q and for EC they were q → p1 and q → p2. For each causal direction, there are four further possible not expected links, for CE: q → p1 and q → p2 and for EC: p 1 → q and p2 → q. If the mean of the causal ratings for any of the not expected links differed significantly from zero that pair of conditionals was excluded. In addition, if for any scenario the mean causal ratings differed between the target relations, that pair of conditionals was excluded. Standard t-tests were used. If the null was rejected for any one of these analyses, that pair of conditionals was excluded. The exclusions left ten scenarios in the causal and seven in the diagnostic condition. Consequently, all materials retained after the pre-test had two unidirectional causal links and each pair was equally sufficient for their effect(s).

The same participants also pre-tested the different materials used in Experiment 2, which are explained below. For experiment 2, six of ten CE conditional pairs were accepted by the pre-test, and six of nine EC pairs.

9.3 Experiment 1: comparison of models

9.3.1 Method

9.3.2 Participants.

40 undergraduate UCL students (5 male, 35 female) took part (mean age: 20.5 years, range: 18 to 25 years). The sample size was determined using prospective Bayesian power analysis (Kruschke, 2013) and the results of Experiment 2 in Ali et al., (2011). A sample size that could lead to similar effect sizes to Ali et al (2011) was sought. In Experiment 2 in Ali et al., the mode of the effect size for the CE-C condition was -2.32 SD units [-3.30, -1.54] ([] = 95% HDI, i.e., Highest Density Interval); for the EC-NC condition it was 2.76 SD units [1.91, 3.71]. The respective means were -.63[-.76, -.50] and .74 [.64, .85]. Simulated data were generated using the modes of the mean and SD for the CE-C condition because the effect size was smaller for that condition. An N of 68 was used to simulate the data, which is the total number of participants in the experiments in Ali et al., (2011). A region of practical equivalence (ROPE) for the effect size was set to .75 SD units. This was set high so as to have sufficient power to detect large effects like those observed in Ali et al., (2011). The analysis showed that a sample size of 40 would provide a .93 (credible interval = .88 to .98) probability that the 95% HDI for the effect size would fall outside the ROPE.

9.3.3 Materials.

The conditionals chosen following the pre-test were embedded in appropriate scenarios and were presented to the participants on PowerPoint slides. Appendix 1 lists the pairs of conditionals that remaining after the pre-test. For example:

"You are meeting a friend in town and know he is planning to drive there. You know that:

If there is an accident on the main road, then he is caught in a traffic jam.

If there are road works on the main road, then he is caught in a traffic jam.

While waiting for your friend to arrive you receive a text message saying he is caught in a traffic jam.

You wonder whether there is an accident on the main road. (Pr(p1|q))

You now remember that there are road works on the main road.

Do you now think it is more, equally or less likely that there is an accident on the main road? (Pr(p1|q,p2)>,=,<Pr(p1|q)?)

You arrive early and so do not know whether or not your friend is caught in a traffic jam.

However, you still wonder whether there is an accident on the main road. (Pr(p))

You now remember there are road works on the main road.

Do you now think it is more, equally or less likely that there is an accident on the main road? (Pr(p1|p2)>,=,<Pr(p)?)

The text in bold is the consequent known condition (C). When the consequent was not known (NC) this text was replaced by the text in italics. As in Experiment 2 in Ali et al., participants were asked for an ordinal change rating of whether p was "more likely", "equally likely", or "less likely" after being told that p2 had occurred (these verbal values were coded as +1, 0, and -1 respectively).

To provide as much variation as possible, within each condition, causal (CE) and diagnostic (EC), participants performed the consequent task (C) with different materials to the not-consequent task (NC).

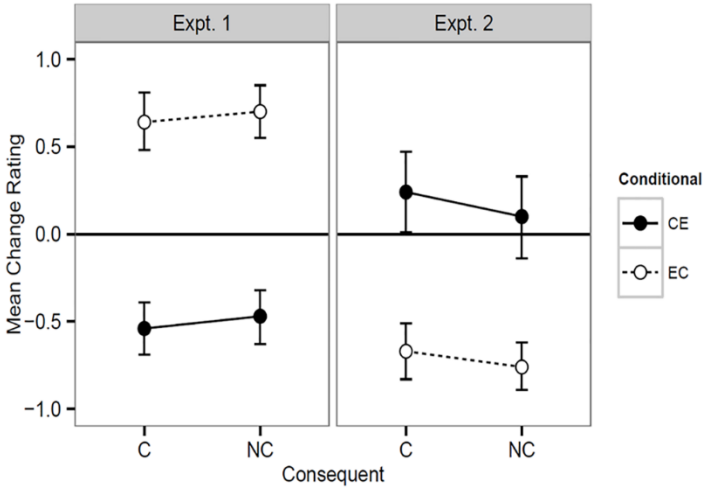
9.3.4 Procedure.

The scenarios were presented randomly to participants on PowerPoint slides. Each scenario was presented on a separate coded slide. A score sheet was also provided with the code of each scenario alongside the three response options: "more likely", "equally likely" and "less likely". Participants circled their response. At the end of the Experiment, participants were thanked for their participation and debriefed as to the purpose of the experiment.

9.4 Experiment 1: Results and Discussion

Figure 4

The results of Experiments 1 and 2



The results of Experiments 1 and 2. Expt. 1 (EC-AND, CE-OR), means and 95% confidence intervals (CIs) based on Model 3 in Table 2. Expt. 2 (EC-OR, CE-AND), means and 95% CIs based on Model 4 in Table 3. The figure is taken from Hall et al., (2016).

Figure 4, Panel A shows the results of Experiment 1. Qualitatively they closely replicated the results of Experiment 2 in Ali et al., (2016) The predictions of augmentation for the EC=NC condition and discounting for the CE=C condition were confirmed. A similar pattern of non-normative responses was also found, that is, augmentation-like behaviour for the EC-C condition and discounting-like behaviour for the CE-NC condition.

The dependent measure, change rating (CR), was an ordinal variable (-1, 0, 1), and the data were analysed using cumulative link mixed models with a probit link function (function clmm in package ordinal implemented in R (Christensen, 2015a; Christensen, 2015b). Participants and items as random effects were added sequentially. Models that corresponded to the predictions of CM (causal models), FM (full mental models), and IM (initial mental models) in Fig 3 were compared. These models are shown in Table 2. Model 1 corresponds to IM in which no effects of causal direction (CE or EC) or consequent (C or NC) are predicted. Model 1 is the null model which predicts the overall mean for all cells in the 2 × 2 design. Model 2 corresponds to CM in which there are main effects of both causal direction (CD) and consequent (C). For both Models 1 and 2, only a random intercept for participants was included. Differences between models were assessed using the likelihood ratio and the Bayes factor. The Bayes factor (BF) is calculated using an approximation based on the Bayesian Information Criterion, BIC, that is, BF=e^((BIC(Model 1) – BIC(Model 2))) (Kass & Rafferty, 1995). Model 2 provided a better fit to the data than did Model 1, $G2(2) = 371.57$, $p < .0001$. The BF in favour of Model 2 is very high. Table 2 also shows the AIC, Akaike Information Criterion, which is another index of fit that does not penalise a model for complexity (the number of parameters) as much as BIC. Model 3 also includes an intercept for items and the BF shows that this model was 5.3×1011 times more likely to have generated the data than Model 2 ($G2(2) = 36.00$, $p < .0001$). Consequently, there were random effects of items that needed to be taken into account. Random slopes for either participant or for items did not improve the fit. This is

illustrated by Model 4 which produced a significantly better fit according to the likelihood ratio ($G2(2) = 8.41$, $p < .001$) but was less likely to have generated the data according to the Bayes factor (i.e., the BF in favour of Model 4 was less than 1). Model 5 corresponds to FM in Fig 2. FM, the fully fleshed out mental model, only predicts a main effect of consequent. Random intercepts were included for participant and item and a random slope for participant because their inclusion led to significantly better fits for Model 4. Model 5 provided significantly poorer fits, $G2(2) = 33.59$, $p < .0001$, and it was 1.7×10-14 times less likely to have generated the data than Model 4.

Table 2

Cumulative link function models for Experiment 1, Hall et al., (2016).

Model	Pars	AIC	BIC	LR	df	BF
1. CR ~ 1 + (1|P)	3	1496.6	1510.2			
2. CR ~ CD + C + (1|P)	5	1129.2	1151.8	371.57	2	
3. CR ~ CD + C + (1|P) + (1|I)	6	1095.2	1122.3	36.00	1	5.3×1011
4. CR ~ CD + C + (1 + CD|P) + (1|I)	8	1090.8	1127.0	8.41	2	0.01
5. CR ~ C + (1 + CD|P) + (1|I)	7	1122.4	1154.0	33.59	1	1.7×10-14

Cumulative link function models for Experiment 1, Hall et al., (2016). Model 1 corresponds to IM (initial mental model); Model 5 corresponds to FM (fully fleshed out mental model); Model 2 corresponds to the CM (causal model) (see Fig 2). Models 2, 3, and 4 correspond to CM with differing assumptions about the structure in the random effects. CR = Change Rating; CD = causal direction; C = consequent. (1|x) = model includes an intercept for the random effect x, either P = participants or I = Items. (1 + CD|P) = model includes an intercept and a slope for causal direction for the random effect of participants. Pars = Number of parameters; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; LR = Log Likelihood Ratio; df = degrees of freedom; BF = Bayes Factor. Each model is compared to the one above it in the list using the likelihood ratio and the Bayes Factor.

The results of these model comparisons showed that a model containing fixed main effects of causal direction and consequent, as predicted by CM, provides better fits to these data than the models predicted by FM and IM. However, Fig 4 also shows that the same pattern of errors occurred as in Experiment 2 in Ali et al., (2016). A model with just a fixed main effect of causal direction actually provides the best fit overall, that is, the predicted main effect of consequent was absent. The inclusion of interactions did not improve these fits. So, in terms of fitting the data, this experiment provides qualified support for CM (Hypothesis 1). Subsequent analyses and the graph in Fig 4 (Experiment 1) are based on Model 3.

To test Hypotheses 2 and 3, the R package lsmeans (Lenth, 2016) was first used to compute asymptotic statistics and significance levels for the simple effects comparisons for Model 3. Next the estimated means and standard errors were used in t-tests for the comparisons to 0. Consistent with CM, but not FM or IM, for the CE-C condition (causal direction, consequent present), the mean change rating was lower than in the EC-C condition, z-ratio = 11.97, $p < .0001$ and less than zero, $t(39) = 7.19$, $d = 2.30$, $p < .0001$. Consistent with CM and FM but not IM, for the EC-NC condition the mean change rating was higher than in the CE-NC condition, z-ratio = 12.02, $p < .0001$ and greater than zero, $t(39) = 9.14$, $d = 2.93$, $p < .0001$. However, consistent only with IM, there were no significant differences between the EC-C and EC-NC conditions, z-ratio = 1.38, $p = 0.17$, nor between the CE-C and CE-NC conditions, z-ratio = 1.39, p = 0.17, although the trends were in the direction (EC-NC > EC-C; CE-NC > CE-C) predicted by CM and FM. However, consistent with none of these theories, the change rating for the CE-NC was less than zero, $t(39) = 5.97$, $d = 1.91$, $p < .0001$, and for the EC-C condition it was greater than zero, p < .0001, and EC-C, $t(39) = 7.69$, $d = 2.46$, $p < .0001$.

These results replicated Experiment 2 in Ali et al., (2016) using an appropriate cumulative link function mixed models approach. The results provide qualified support for causal model theory because it is the only theory that predicts a fixed main effect of causal direction, which was the dominant finding. In so far as neither FM nor IM predict the main effect of causal direction these results argue against the mental models account.

As observed, the data only offered qualified support for CM. Moreover, as mentioned in the introduction an alternative shallow encoding hypothesis may also explain this result. Experiment 2 was designed to provide a test of this hypothesis.

9.5 Experiment 2: Shallow Encoding

9.5.1 The rationale for this experiment: an explanation for non-normative discounting and augmenting in the change mode

Experiment 1 of Hall et al., (2016) replicated the surprising finding of Experiment 2 of Ali et al., (2011). When asked to give a directional description of the effect of learning the truth of one antecedent on belief in the likelihood of the other in conditional pairs, participants discounted when it was appropriate for common effect conditional pairs (i.e., when the consequent was known to be true), but also when it was not appropriate (i.e., when the status of the consequent was not known). For common cause conditional pairs, participants augmented when appropriate (i.e., when the truth of the consequent was not known), but also when augmenting was not appropriate (when the consequent was known to be true).

A possible explanation of this pattern of errors for the change mode, due to Mike Oaksford, concerns the requirements on participant reasoning in the change rating response format.

Here are two example pairs of conditionals from the materials used in Experiment 1:

Common-effect (CE)

1 If the fuse on the stereo is blown, then the stereo is off (if p1, q)

If the stereo is unplugged, then the stereo is off (if p2, q)

Common-cause (EC)

2 If she feels dizzy, then she is hungry (if p1, q)

If her stomach is rumbling, then she is hungry (if p2, q)

Although the model comparison of the results of Experiment 1 favoured a causal model, the support for that theory was good relative to the models based on mental model theory, but not impressive in an absolute sense: results showing no effect of the consequent are not consistent with causal models. Experiment 2 aimed to test a hypothesis compatible with these results, termed here 'shallow encoding'. Taking the CE sentences above, participants are told that the stereo is off and then that the stereo is unplugged (CE-C) or just that the fuse is blown (CE-NC). In both conditions, they were asked whether the stereo being unplugged increased or decreased the likelihood of the fuse being blown or left it the same. This response format imposes a memory load. In the CE-C condition, for example, participants must interrogate their mental representation assuming the stereo is off and they have no information about the stereo being unplugged and assess the probability that the fuse has blown. They then must interrogate the model again assuming the stereo is unplugged and re-compute the probability that the fuse has blown. They must then compare this new result with the result they had to retain in working memory from the previous query. Earlier research has suggested even small memory loads like those imposed by easily visualizable materials can disrupt reasoning (e.g., Knauff, 2013).

The simple strategy proposed to explain the results of the change mode in Experiment 1, and in Experiment 2 of Ali et al., (2011) is inspired by mental models theory, which suggests that, due to, for example, working memory limitations, people frequently only construct partial representations of logical relations called "initial models", as described in the introduction. The results of Experiment 1 seem to indicate that people ignore the consequent manipulation. One reason for this could be that, because of the additional memory load imposed by the response format, they only formulate a partial representation of the CE and EC conditional pairs which excludes the consequent. So for the CE pair, which is recoded as *if the fuse on the stereo is blown* (*p1*) OR *the stereo is unplugged* (*p2*), *then the lights go out* (*q*), the partial mental model $p1 \quad \neg p2$, $\neg p1 \quad p2$ is constructed representing the possibilities allowed by an exclusive-or between *p1* and p2. On this reading, if *p1* occurs *p2* should not and if *p2* occurs *p1* should not. For the EC pair, which is recoded as *if someone is hungry* (*q*), *then she feels dizzy* (*p1*) AND *her stomach is rumbling* (*p2*), the partial mental model $p1 \quad p2$ is constructed representing the only possibility allowed by a conjunction between *p1* and *p2*. On this reading, if *p1* occurs so should p2 and vice versa. This "shallow encoding" hypothesis predicts discounting in both the CE-C *and* the CE-NC conditions and augmenting in both the EC-NC *and* the EC-C conditions, which is what was found in Experiment 1, and for the change mode in Ali et al., (2011).

Shallow encoding based on mental models is compatible with both mental models and a probabilistic approach based on CBNs. It has recently been argued that mental models might provide the right kind of representation for recording the results of interrogating or sampling underlying probabilistic representations and that this might explain certain errors as people move from continuous to discrete representational formats (Hattori, 2016; Oaksford & Hall, 2016). This proposal means that some representational format like mental models would be required even if one takes a probabilistic view of the underlying deep logical structures over which people normally reason.

The shallow encoding hypothesis can be tested by using materials which suggest a relationship between the two conditionals that runs counter to that described above. That is, pairs of causal conditionals (CE) that can be recoded as *if p1 AND p2, then q* and pairs of diagnostic conditionals (EC) that can be recoded as *if q, then p1 OR p2*. If the shallow encoding hypothesis is true, then the results for the causal and diagnostic conditionals should be mirror images of the results of Experiment 1, and Experiment 2 in Ali et al., (2011). That is to say, augmentation-like behaviour for the causal conditionals (CE) and discounting-like behaviour for the diagnostic conditionals (EC) independent of the status of the consequent.

Materials that implement this manipulation are (6) for the conjunctive antecedents and (7) for the disjunctive antecedents:

If the plant is watered often (p1), then it grows well	(q)	6
If the plant receives light (p2), then it grows well	(q)	

If the participant wears glasses (p1), then their vision is poor	(q)	7
If the participant wears contact lenses (p2), then their vision is poor	(q)	

However, apparently only materials like (6) are likely to prove discriminatory. (7) achieves an appropriate exclusive-OR reading, i.e., *if a participant's vision is poor, then they wear glasses OR contact lenses BUT NOT BOTH.* However, any implementation in a CBN would lead to the same conclusion that people should discount whether it is known the person's vision is poor or not. The simplest implementation would be a two-node network $q \rightarrow C$, where C is a three level corrected vision variable with three mutually exclusive levels, *glasses* (p), *contacts* (r), and *nothing*. Such an implementation would make the same predictions as the shallow encoding hypothesis.

The materials in (6), however, could be discriminatory. Section 5.5 gives a formula for a noisy-AND integration rule, where both causes are necessary for the effect.

Using this rule in the common-effect structure predicts augmentation when it is known that the plant grows well (CE-C) but not when this is not known (CE-NC). Knowing that the plant grows well and is watered often increases the probability that it also received light. Augmentation has been observed in a very similar condition (Rehder, 2015). Consequently, for this pair of conditionals, the predictions based on the noisy-AND rule contrast with the shallow encoding hypothesis in predicting no augmentation for the CE-NC condition.

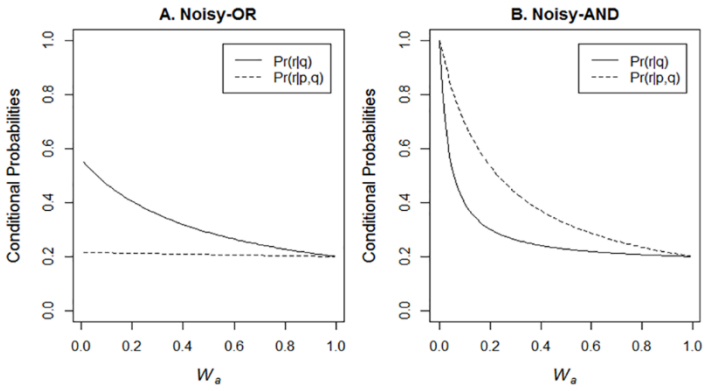
The contrastive predictions are thus that noisy-AND predicts less augmentation for the CE-NC condition in this experiment than for the EC-NC condition in Experiment 1 but augmentation should be at similar levels for the CE-C condition as for the EC-C condition in Experiment 1. However, the noisy-AND formula shows that the probability of alternative causes could allow the CE-C condition to also discriminate between noisy-AND and the shallow encoding hypothesis. Figure 5 compares the noisy-OR and noisy-AND integration rules, and Fig 5B shows that augmentation reduces as W_a approaches either 0 or 1. At either extreme, no augmentation is predicted. A manipulation that should reduce the scope for alternative causes is when *p1* and *p2* are individually necessary for the effect. For example, plants will not grow well in the absence of water or light, that is, there is no alternative cause that can make plants grow well in the absence of these necessary causes. Using materials which encourage this interpretation should have the effect of reducing augmentation for the CE-C condition. If this happens, then augmentation should be lower in a CE-C condition in this experiment than in the EC-C condition in Experiment 1.

1. According to both CM and shallow encoding, in the diagnostic direction (EC) change ratings should be lower than in the causal direction (CE) and below zero, that is, discounting-like behaviour should be observed.

2. According to CM, dependent on the manipulation of causal necessity, in the causal direction (CE), the change ratings should not differ from 0 for CE-NC but there should be some augmentation for CE-C, that is, the change rating should be greater than 0.
3. Again dependent on the causal necessity manipulation, the levels of augmentation for CE in Experiment 2 should be much lower than for EC in Experiment 1 for both levels of consequent.

Figure 5

Discounting with the noisy-OR and augmentation with the noisy-AND integration rules



Discounting with the noisy-OR (A) and augmentation with the noisy-AND (B) integration functions varying the probability of alternative causes Wa. For noisy-OR, unless W_a=1, *Pr*(r|p,q)<*Pr*(r|q) and discounting is predicted. For noisy-AND, unless W_a= 1 or 0, *Pr*(r|p,q)>*Pr*(r|q) and augmentation is predicted. In these graphs *Pr*(p)=*Pr*(r)= .2 and Wp=Wr=Wpr= .9. Figure taken from Hall et al., (2016).

9.5.2 Pre-test

For details of the pre-testing of materials, see section 9.2.1

9.5.3 Participants.

The sample size for the main experimental phase was smaller (N = 28) than in Experiment 1 but it was drawn from the same population, i.e., students at University College London. Sample size was set from a prospective power analysis as in Experiment 1 (Kruschke, 2013). CM predicts no (CE-NC) or less (CE-C) augmentation for causal conditionals. According to shallow encoding, augmentation should be of a similar magnitude that that observed in Experiment 1 for the diagnostic/not-consequent condition. Consequently, a sample size was sought which could allow effect sizes of similar magnitude to Experiment 1 to be observed. Therefore, as in Experiment 1, the ROPE on the effect size was again set at .75 SD units. Simulated data were generated using the mean and SD for the diagnostic/not-consequent condition in Experiment 1 with an N of 108 which corresponds to the total number of participants in Ali et al., (2011) and Experiment 1 where augmenting was observed. The prospective power analysis indicated that a sample size of 25 would provide a .93 probability that the 95% HDI for the effect size falls outside a .75 SD ROPE, i.e., a .93 probability that the null can be rejected. Consequently, an N of 28 provides sufficient power to detect an augmentation effect of this magnitude for the causal conditionals if there is one.

9.5.4 Materials.

The conditional pairs used for this experiment are listed in Appendix 1. In this experiment, for the CE-AND conditions materials were used to encourage an interpretation were the causes were individually necessary for their effects. This was to encourage a reduction in augmentation for the CE-C condition (Rehder, 2015). Confirmation that the CE rule pairs were interpreted as individually necessary came from an additional on-line test using another 47 participants and two subsets of these materials. Ratings were collected for Pr(q|p), Pr(q|r), Pr(q|¬p) and Pr(q|¬r) for these rule pairs in the CE conditions. Pr(q|¬p) and Pr(q|¬r) were subtracted from 1 and averaged, to reveal differences in probability. In the CE-AND condition

in this experiment (*Pr*(q|x): $\bar{m}$ = .77, SE = .03; *Pr*(¬q|¬x): $\bar{m}$ = .69, SE = .03) causes were rated as less sufficient, t(13.33) = 2.99, p < .01, and more necessary, t(13.67) = 2.86, p < .01, than in the CE-OR condition (*Pr*(q|x): $\bar{m}$ = .89, SE = .03; *Pr*(¬q|¬x): $\bar{m}$ = .57, SE = .30) in Experiment 1. For three rule pairs the cause was rated as more necessary than sufficient. The results were analysed in three ways, (i) using all the materials, (ii) restricting the CE-AND condition to just Rule pairs 2, 4, 5, and (iii) using just the subset of rule pairs for which conditional probability ratings were collected plus the restriction in (ii).

9.6 Experiment 2: Results and Discussion

Figure 4 Panel B shows the results of Experiment 2. As for Experiment 1, qualitatively they closely follow the pattern predicted by CM. That is, in the CE causal direction there was no augmentation for CE-NC but some for CE-C. Consistent with the shallow encoding and CM, in the EC direction there was discounting, regardless of whether the consequent was present (C) or absent (NC). The result was analysed in the three ways mentioned at the end of the section Materials. However, which subsets of materials were used made no difference to the results. Consequently, unless stated the results using all materials are reported.

Table 3

Cumulative link function models for Experiment 2

Model	Pars	AIC	BIC	LR	df	BF
1. CR ~ 1 + (1|P)	3	710.7	722.0			
2. CR ~ CD + (1|P)	4	597.3	612.6	117.30	1	
3. CR ~ CD + (1|P) + (1|I)	5	591.7	610.8	7.64	1	6.17
4. CR ~ CD + C + (1|P) + (1|I)	6	591.7	614.6	1.94	1	0.02

Cumulative link function models for Experiment 2. Model 1 is the baseline no effect model (i.e., the overall mean); Models 2, 3, and 4 correspond to CM with differing assumptions about the structure in the random effects. All acronyms are the same as in Table 2. Each model is compared to the one above it in the list using the likelihood ratio and the Bayes Factor.

The cumulative link function models which were compared are shown in Table 3. Model 1 is the null model. Model 2 corresponds to a model in which there is only a main effect of causal direction (CD). This model provides a better fit than the null model, $G2(1) = 115.25$, $p < .0001$. Adding a random effect of items, Model 3, improves the fit, $G2(1) = 7.64$, $p < .01$, and it is 6.17 times more likely to have generated the data than Model 2. Model 4 shows that adding a main effect of consequent (C) does not improve this fit. The fits were not improved by including either interactions (fixed effects) or slopes (random effects). According to the AIC, adding a main effect of consequent did not worsen the fit either. We, therefore, used Model 4 for the means shown in Fig 4 (Expt. 2) that were compared to test Hypotheses 4 and 5.

Consistent with CM and the shallow encoding hypothesis, the mean change ratings for the EC-C condition were lower than for the CE-C condition, z-ratio = 7.05, $p < .0001$. The

mean change ratings for the EC-NC condition were also lower than for the CE-NC condition, z-ratio = 6.80, $p < .0001$. The change ratings were less than zero for the EC-C condition, $t(27) = 8.08$, $d = 3.11$, $p < .0001$, and for the EC-NC condition, $t(27) = 10.72$, $d = 4.13$, $p < .0001$. These results confirm Hypothesis 4. The change rating for the CE-C condition was greater than zero, $t(27) = 2.05$, $d = .79$, $p < .025$ (one-tailed), but it did not differ from zero in the CE-NC condition, $t(27) = .82$, $d = .32$, $p = .42$. The Bonferroni correction for these multiple hypothesis tests mean that the significance level required for each individual hypothesis is $.05/2 = .025$. So the result for the CE-C condition was just significant. However, the effect size was in the medium range ($d = .79$) and fell outside the ROPE ($d = .75$) that was set in the power analysis to detect an augmentation effect. This was in contrast to the CE-NC condition which fell inside the ROPE. So some confidence can be placed in this result for the CE-C condition. Consequently, using materials for which the causes were individually necessary for the effect has reduced but not eliminated augmentation for the CE-C condition in this experiment, confirming Hypothesis 5.

To test Hypothesis 6 the levels of augmentation for the CE causal direction in this experiment were contrasted with those observed for the EC direction in Experiment 1. The best fitting cumulative link function model had fixed main effects for causal direction and consequent with random intercepts for participants and items. This newly estimated model produced minor variations in the means for the fixed effects but the results of the simple effects comparisons were clear. The change ratings were a lot lower for the CE-C condition in Experiment 2 than for the EC-C condition in Experiment 1, z-ratio = 5.04, $p < .0001$. They were also a lot lower for the CE-NC condition in Experiment 2 than for the EC-NC condition in Experiment 1, z-ratio = 5.01, $p < .0001$. These findings are not consistent with the shallow encoding hypothesis. This hypothesis must predict that the levels of augmentation should be the same in both conditions because the recoding of both pairs of rules involves a conjunction.

The original report of these findings, in Hall et al., (2016), further considered whether recent developments in mental model theory (Johnson & Byrne, 2002; Johnson et al., 2015). That discussion is not repeated here, inasmuch as 'new MMT' with the added-on possibility of 'pragmatic modulation' to take account of common-sense reasoning, remains an obscure and moving target. Hall et al., (2016) had to make informed guesses as to what new MMT might predict. Here I follow the conclusion of Cruz (2018, p. 96) that, until it is elaborated and clarified, new MMT remains 'not worth pursuing'. The original discussion can be found in Hall et al., (2016). Here I will conclude that the results of these two experiments do not support mental model theory.

9.7 CBNs and these results

Ali et al., (2010), Ali et al., (2011), and Hall et al., (2016), seem to confirm that an account of how people reason with conditional statements needs a central place for marking causal direction. Further, these three papers subjected the two main competing accounts of conditional reasoning that are not logic-based, MMT and causal reasoning, to extensive comparisons to examine their ability to model conditional reasoning. MMT's core account is, like logic, not able to account for the common-sense, intuitive, and empirically demonstrated effects of discounting and augmentation. Attempts in recent years to adapt MMT to handle these phenomena, and others, seem ad-hoc and less than coherent, and the research presented in these three papers supports the claim that only a probabilistic approach offers a rich enough framework to handle causal conditional reasoning, and at the same time, to allow causal conditional reasoning to be integrated into a more general understanding of reasoning and language. Currently, the CBN formalism, widely used in real-world, non-psychological, applications, seems to be the most productive framework for a computational-level account, at least, of how people use information about causes and effects. Further, since CBNs provide an

effective way to carry out counterfactual reasoning, it seems helpful to model even non-causal conditional statements using this approach.

At the same time, these experimental results leave the response mode discrepancy unexplained by MMT, CM, and also by an explanation for the change mode results based on simple associative reasoning.

Chapter 10 This research

10.1 The rationale

The experiments to be reported here were carried out with two aims.

Firstly, to see if the results of Ali et al., (2010), Ali et al., (2011), and also of Hall et al., (2016) are robust and replicable. These experiments found discounting and augmenting in reasoning about conditional scenarios. These are phenomena which are predicted when conditionals are understood according to their causal content, rather than according to their logical form only. Traditional logic, and the basic MMT approach, see the meaning of conditional sentences as deriving from their component clauses as antecedent and consequent. A causal understanding sees the meaning of a causal conditional as determined by the cause and effect referred to by its clauses. Causation is considered as a basic property of either the world itself, or at least of human reasoning when predicting the world. Pairs of conditionals sharing a cause, or sharing an effect, describe 3-node causal networks, for which discounting and augmenting are normative. However, no experiment in the earlier research found normative judgements in respect of discounting and augmenting behaviour in all conditions. The results are consistent neither with reasoners representing conditional sentences as symbols to be manipulated according to rules, nor with reasoners working with conditionals as representations of causal relations in the real world. Further, the earlier research found, in separate experiments, an unexpected discrepancy between results for the two response modes used. For this response mode discrepancy, whether these findings can be replicated is an aim of the experiments reported here. Finally, one of the response modes (the delta mode) itself so far lacks a successful replication, inasmuch as Ali et al., (2010) found normative behaviour for the EC-C condition (in children), while for the same condition Ali et al., (2011), found non-normative discounting-like behaviour with adult participants. This research uses adult

participants only. The replications differ in two ways, considered to be refinements: response mode is a within-subjects factor, a more powerful procedure, and the experiments were presented via the internet to participants recruited online.

Secondly, the experiments carried out were extended beyond the earlier research by collecting additional data, which it was hoped would be helpful in supporting, or ruling out, possible explanations for such discounting and augmenting reasoning behaviour as might be found. Data relating to conditional probabilities, and participants' understanding of the world beyond the presented conditionals may give an explanation for the way in which people discount and augment. Data relating to individual differences between participants may give an explanation for the response mode discrepancy, if found. The assumption here is that, since the causal (and logical) content of the conditionals to be presented will be the same across the response modes, a discrepancy will not be due to that causal content. At the same time, given that the within-subjects design of the replication meant that the same scenarios were presented to participants in both delta and change response modes, the collection of explanatory data, important in itself, also served as ‘filler’, collected mostly in experimental phases that separated the response modes. The aim here was to reduce any tendency for participants to be influenced by a response to one mode when giving a judgement in the other. At the beginning of the series of experiments, this exploratory data was of two types. A working memory task was given to participants. This was because of the presumably greater WM load associated with the change mode. Unlike the delta mode, where the comparison of the judgements made before and after the assertion of one of the consequents was made using the collected data, by subtraction, in the change mode the comparison was a task assigned to the participants, requiring one value to be held in memory while the other was assessed, before a comparison was made. The other data collected were a series of judgements of conditional probabilities (and a joint probability) relating to the clauses of the conditionals used in the scenarios. The conditionals presented

were built around actions and states known to participants from the real world. Participants' pre-existing beliefs about the real world must affect how strong, or plausible, they see the causal relationships suggested to them as being, and will presumably influence their judgements. These data were collected with the aim of assessing that influence.

These dual aims will form the basis of the reporting of the results, with the status of attempted replication reported before, and separately from, that data collected which has no counterpart in the earlier research, and which is intended to be explicatory of any replication. Reporting the results by analysis rather than by the chronological sequence of the experiments is also necessary and appropriate inasmuch as some of the analyses of the exploratory data could not, due to insufficient sample sizes, be carried out for each experiment separately. For example, the analysis of the WM data, by participant WM score, required combining the data from all experiments to give sufficient data to allow the MCMC chains of the Bayesian models which were fit to converge.

10.2 Overview of the experiments

Six replicatory experiments were carried out, plus an additional experiment which deliberately diverged from replicating Ali et al., (2010), Ali et al., (2011), and Hall et al., (2016) with the aim of seeing if a manipulation of the materials and procedure could cause a failure to replicate. The six replication attempts formed three pairs, each with a repeat experiment using a different web platform. For those parts of the experiments relating to the replication of the earlier research, the only difference was that for the first experiment, and its repeat, participants, as in Ali et al., (2011), and Hall et al., (2016), gave their judgements by supplying an integer value (0-10). In the other experiments, participants responded by dragging an on-screen slider control to indicate their judgement. This gave a more finely-grained response. This change in the replication portion of the experiments was made in part to keep the response method in line

with the conditional judgements. Calculations of the phi coefficient made from these responses in the first experiment, and its replication, produced too many cases of division by zero when only 11 response values were available. In the third pair of experiments, an attempt was made to collect timings of how long participants took to answer. This was done in a manner that was transparent to the participants. These experiments also introduced questions for the participants on how confident they felt about their responses. This change in materials and procedure made this pair of experiments appear somewhat different to the participants, in as much as new questions had been added. Other changes in the sections of the experiments in which data for analyses exploratory of possible explanations for discounting and augmenting effects were small improvements and corrections to shortcomings discovered in the materials.

10.2.1 The experimental template

Thus, all seven experiments were based upon the first, with small changes introduced in light of experience, but only the final experiment (experiment 4) was deliberately intended to deviate from the first.

For clarity an overall experimental template, a description of the experimental design for the first experiment carried out, and a brief summary of the materials used, are first presented here, differences from which will be noted below when necessary, as the remaining six experiments are introduced. The most important of these changes will be (i) a change from those cases where judgements are indicated by entering an integer to using a slider-type control, (ii) a change in how the conjunctive probability is collected, to increase the compatibility of participants' ratings with the axioms of probability, and (iii) a correction to the wording of the conditional ratings for the EC condition. A fuller description of experiment 4, which differed in rationale from the others, is given at the end of this section.

10.2.2 The four phases of the experiments

The first experiment was made up of four phases (as were all subsequent experiments). The scenarios to elicit discounting and augmentation were presented in the first and fourth phases. Each scenario was presented first in phase one, and later presented once again, but in the other response mode, in phase four. The second and third phases thus provided distracting filler material for participants between the two inferential phases. Since they saw the same scenarios twice in this within-subjects design, it was hoped that when responding in phase four participants would not be influenced by their response from phase one. At the same time, such influence was less likely because no participant saw a particular scenario twice with the wording differing only in whether the consequent was asserted. This is because phases one and four differed by response mode – a scenario appearing in its delta version in one of the phases would be seen in the change version in the other phase. Whether a participant saw the change or the delta version first depended on their random assignment to one of two experimental paths through the material. Phase two was a test of working memory ability. In the third phase, for each scenario, seven questions elicited participants' judgements of conditional and conjunctive probabilities associated with the clauses of the conditionals for that scenario. Within each of the four phases, tasks were presented in random order.

10.3 An overview of the experimental procedure

Details are given here of a basic template for the experiments, deviations from the which will be noted in the results as necessary. The experiments fell into four separate phases, of which the first and the fourth were attempted replications of the earlier research (Ali et al., 2010; Ali et al., 2011; Hall et al., 2016).

10.3.1 Phases 1 and 4

The first and fourth phases presented 24 scenarios in a 2 × 2 × 2 design, with causality (CE: common effect /EC: common cause), both of which were determined by the scenario, and presence of consequent (C: consequent asserted /NC: consequent not asserted) as between-subject factors, and response mode (change rating/delta rating) as a within-subjects factor. Participants were randomly assigned to one of two routes through the experiment (the variable recording the particular route is referred to below as 'blocks'). In the first route, the first phase used the delta response mode and the fourth used the change response mode. In the second route the response mode was reversed between phase one and four. The routes also differed in consequent condition used (C or NC) - those scenarios which were shown with the consequent present (absent) in route 1 were shown with the consequent absent (present) in route 2. This splitting of the main experimental stimuli was done to reduce and reveal effects of having already answered one assessment type on responses to the other. (However, analysis of the results showed no effect of order.)

The dependent variable for the change ratings mode of the scenarios in the first and fourth phases was a choice of three values ('Less likely' / 'Equally likely' / 'More likely'), and for the delta versions, 0 to 10 ('certainly not' to 'certainly') ratings of one antecedent (P1) before and after assertion of the other antecedent (P2).

The dependent variable for the difference scores mode of the scenarios in the first and fourth phases was calculated from integer ratings between 0 and 10 typed in by participants as values for 'how likely' P1 was both before and after assertion of the other antecedent (P2). The difference score was the value of 'rating after' minus 'rating before'.

10.3.2 Phase 2

The second phase was a simple working memory test. Participants were asked to deduce a sequence of 4 letters on the basis of 3 previously, separately, presented statements about the

order of particular pairs of letters. There were 16 questions presented in random order. The dependent variables were accuracy in choosing the sequence, and the time spent viewing the answer screens before responding. This task is described in Woltz (1988), with the difference that in the current experiment no feedback on performance was given to participants, and in Kyllonen and Christal (1990). It is a test of both deductive ability and working memory, measured by both accuracy and speed. (See materials section below for example tasks from two of the experiments).

10.3.3 Phase 3

The third phase asked for 7 probability ratings relating to each of the scenarios (6 conditional probabilities and 1 joint probability) in the form of their likelihood on a 7-point scale. These questions were presented in random order.

10.4 An overview of the materials

Next follows an overview of the materials used for the four experimental phases described above. These are taken from one CE scenario, 'vase', except for those relating to the WM task, which are not related to a particular scenario.

10.4.1 Example of the materials relating to one causal scenario

Phases 1 and 4:

The change mode ('C', consequent asserted, version):

You have recently moved into a new flat. One morning you receive a package from your aunt. You open it up and see that she has given you and your friend a vase as a housewarming gift. Your friend thinks that this vase is especially hideous and tells you the following information:
If she throws the vase, then the vase breaks.
If a tennis ball hits the vase, then the vase breaks.

A few days later you are out at work and while on the phone to your flatmate, you hear in the background that the vase breaks.

You wonder whether she throws the vase.

She tells you that a tennis ball hits the vase.

Do you now think it is less, equally or more likely that she throws the vase?

Less likely | Equally likely | More likely

The delta mode:

You have recently moved into a new flat. One morning you receive a package from your aunt. You open it up and see that she has given you and your friend a vase as a housewarming gift. Your friend thinks that this vase is especially hideous and tells you the following information:

If she throws the vase, then the vase breaks.

If a tennis ball hits the vase, then the vase breaks.

A few days later you are out at work and while on the phone to your flatmate, you hear in the background that the vase breaks.

You wonder whether she throws the vase.

How likely do you think it is that she throws the vase?

Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

[new screen]

She tells you that a tennis ball hits the vase.

How likely do you now think it is that she throws the vase?

Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

(An example of each scenario is given in appendix 4.)

Phase 2:

An example question for the response DCBA (the terms ‘Set 1’ and ‘Set 2’ were explained in an example question as referring to letters in positions 1 and 2, and in positions 3 and 4, respectively):

Set 1 presented first until dismissed:

A is not followed by B

Set 2 presented second until dismissed:

C is preceded by D

Set order presented third until dismissed:

Set 2 does not follow Set 1

The question:

Which of the following eight items did the previous three statements describe?

With a list of the choices ABCD, ABDC, BACD, BADC, CDAB, CDBA, DCAB, DCBA

Phase 3:

One of the 6 conditional probability questions, and the joint probability question for the vase scenario:

Pr(Q given P1)

If she throws the vase, then the vase breaks.

If a tennis ball hits the vase, then the vase breaks.

Suppose that she throws the vase.

How likely is it that the vase breaks?

Pr(P1 and P2)

If she throws the vase, then the vase breaks.

If a tennis ball hits the vase, then the vase breaks.

How likely is it that she throws the vase AND a tennis ball hits the vase?

In each case the participant responded by clicking one of the answers from this ordered list:

'Definitely not', 'Highly unlikely', 'Unlikely', 'Perhaps / Perhaps not', 'Likely', 'Highly likely', 'Definitely yes'

In what follows the seven experiments are each denoted by a number and a letter. The number, 1, 2, 3, and 4, refer to the four slightly differing experimental designs, the first three of which were separately repeated using a different web platform. Experiment 1A and its repeat experiment, 1B, are the closest attempted replication of Ali et al., (2011) and experiment 1 in Hall et al., (2016). As noted above, experiments 2A and 2B differ as replications by a change from asking participants to give their judgements in the delta mode by entering an integer, to responding by dragging an on-screen slider to a position indicating their judgement. Experiments 3A and 3B differ further as replications as questions about participants confidence in their answers were included just after their responses in the delta and change modes. Experiment 4 included changes which made it no longer a replication of the earlier research. These changes were made to examine whether the response mode discrepancy might be to some extent due to the wording and presentation of the stimuli having led to a failure by participants to fully consider their task in the change mode. Thus the experimental design progressed in the order 1A and 1B, 2A and 2B, 3A and 3B, and finally 4. On the other hand, chronologically the experiments were carried out in the order 1A, 2A, 3A, 1B, 2B, 3B and finally 4.

10.4.2 Experiment 4 – how and why this experiment differed from the others

For experiment 4, significant changes to the wording and presentation of the scenarios were made which made the sequence (e.g., two screens for both scenarios, rather than two screens for the delta mode, and only one for the change mode) less different between the two modes. This was done to see whether such a change in presentation, but not in scenario content,

might make the results for the change mode more similar to the those for the delta mode. The intention was to reduce as far as possible differences between the modes not related to the difference in causal direction. Only these particular revisions for this experiment alone are described here. Other changes than these, for all experiments other than 1A, are described along with the relevant results for those experiments.

The results from the change response mode are consistent with participants not taking account of the assertion or non-assertion of the consequent. On the other hand, participants have apparently taken notice of which of the two normative behaviours, discounting and augmentation, would be appropriate given the causal direction of the scenarios. That is to say, for predictive conditional pairs, with two causes and a single effect, participants discounted (responded 'less likely') where it would be normative, but also did so in the inappropriate (consequent absent) scenarios. Similarly, for diagnostic conditional pairs, with two effects of a single cause, participants augmented (responded 'more likely') where appropriate, but also did so where it was not normative (in the consequent absent cases).

All the information for the process of assessment and updating was presented in the change mode to participants together in a single screen, while in the delta mode the consequent, where asserted, appeared on the first screen, separated from the assertion of the second antecedent, which was on a second screen, which participants called up by actively clicking to dismiss the first. It was considered possible that this lack of a clear temporal progression in the change mode may have made it less easy for the participants to realise that an updating process was required. The pattern of results is consistent with participants in the change mode not carrying out a process of belief revision on learning of the second antecedent, for which probabilities would depend upon the status of the consequent. Rather, the answers are consistent with answering the question 'without considering what is known is this case, is this a scenario where the effect of one antecedent given the other would be make more or less likely

by knowledge about the consequent?' It was thought possible that the progression across two screens, as used in all experiments for the delta mode, might help participants to realise updating of their judgements was required.

The materials were altered to make the sequence of the two modes more similar, by requiring participants in the change mode to also dismiss the first screen in order to progress to a second screen, where the second antecedent was now asserted, and judgement asked for. Further, added wording was added to both modes to make the required updating process more clear. This change in wording was expected to have no effect on the delta response mode, where the two judgements in the updating process are already distinct. Taken together, these changes aimed to clarify whether the differences in materials, rather than some inherent difference in participants' causal reasoning strategy, are responsible for the response mode discrepancy.

The changes in wording were as follows, taking the CE 'vase' scenario as an example:

The statement of the second antecedent was introduced with the word 'Next,'

Next, she tells you that a tennis ball hits the vase.

The query of the judgement (change) or second judgement (R2 delta R2) was introduced with a statement of the accumulated information in light of which that judgement was to be made, e.g. for delta (C)

The vase breaks, and you have just learned that a tennis ball hits the vase, how likely do you now think it is that she throws the vase?

and for change (C)

The vase breaks, and you have just learned that a tennis ball hits the vase, do you now think it is less, equally or more likely that she throws the vase?

For the NC condition this query began at the phrase 'You have just learned...'

10.5 An overview of the statistical analyses performed

The statistical analyses used on the data from these experiments are a third application of the Bayesian approach. CBNs, described above, are Bayesian. The probabilistic approach to reasoning, also introduced above, is Bayesian. The use of Bayesian statistics to analyse data is an alternative to traditional methods, often called 'null hypothesis significance testing' (NHST). This, modern, statistical approach is independently motivated from the other two Bayesian applications introduced so far, and has its roots outside psychology. Bayesian data analysis is used more widely, and was used earlier, in fields unrelated to psychology and reasoning, and traditional frequentist (null hypothesis significance testing: NHST) analyses are still at present dominant in research within the new paradigm of reasoning psychology, and Bayesian analyses are not more appropriate for data related to reasoning tasks as opposed to other data. Nonetheless, the arguments in favour of Bayesian, rather than frequentist, analysis of experimental data are compelling (Kruschke, 2014). A Bayesian approach also promises a more coherent and thoroughgoing approach to freeing psychology research from its 'replication crisis' beyond the use of confidence intervals and effect sizes (Cumming, 2013).

Cummins et al. (1991) demonstrated that experimental participants reasoned differently about causal situations as a result of their pre-existing understanding of the world. Therefore, to isolate the effect of the assertion of the consequent from pre-existing beliefs about the world due to the twelve different scenarios presented, the data here are analysed using a multi-level regression, in which the scenario presented was a random factor. Using a Bayesian analysis, the central measures of the parameters for each scenario were modelled as themselves coming from a higher parameter distribution, that for the consequent present / absent fixed effect factor. Similarly, participant was modelled as a random effect, subordinate to the fixed effects of consequent and causal direction. Thus, these analyses use maximal random effects models (Barr, Levy, Scheepers, & Tily, 2013). These analyses were carried out using the brms package

(Bürkner, 2017, 2018), which is an interface to the Stan (Carpenter et al., 2017) package which does not require models to be specified in the Stan language. A particular convenience of brms is that it enables both multi-level regression and multi-level ordinal regression models, appropriate for the delta and change response modes respectively.

For all the analyses using Stan presented here, the default priors chosen by the brms package were accepted without change. For the fixed (population-level) effect brms assigns improper flat priors (Bürkner, 2017).

Visual inspection of the sampling chains suggested good convergence of the model, and this was confirmed by R-hat diagnostic statistics, which were less than 1.1 for all the main effects (Gelman et al., 2013) except as otherwise noted.

Models were compared by means of a leave-one-out (LOO) information criterion (LOOIC), (Vehtari, Gelman, & Gabry, 2017), calculated using the R package LOO, (Vehtari, Gabry, Yao, & Gelman, 2018).

The posterior distribution of the model coefficients was sampled to obtain a distribution of response probabilities. Stan estimated coefficients from the multi-level model as distributions (rather than point values) for three types of uncertainty: at the population level (fixed effects), at the group level for scenario and participant (random effects), and residual variance (uncertainty in the observations not otherwise accounted for by the model). The model response values were sampled using the 'fitted' function supplied by brms, called so as to not take account of the group-level (random) effects, thus giving posterior draws taking into account only the highest-level effects of consequent and causal direction. This same procedure was followed for all experiments.

10.6 Participant summary

All seven experiments were prepared on the Qualtrics web platform (https://www.qualtrics.com/uk/), and were presented to paid participants recruited over the internet via either Crowdflower (https://www.figure-eight.com/ : the company has rebranded since these experiments were conducted) or Prolific Academic (https://prolific.ac/).

For Hall et al., (2016), a Bayesian analysis was carried out, simulating data with effect sizes matching those of Ali et al., (2011), to determine a suitable sample size. Inasmuch as phases 1 and 4 of the present experiments were close replications of the earlier research, similar sample sizes were used in this research. The power analysis in Hall et al., (2016), gave a sample size of 40 as likely to be sufficient to confirm the effects sought. As participants were recruited over the internet, slightly larger participant requests were made, to allow for some attrition due to failures to complete, and elimination of some participants due to completion in suspiciously short times, or possible inclusion of participants who had taken part in previous experiments. In each case, the analysed sample size was more than 40.

Participant recruitment was restricted to native English speakers, and those who had not completed earlier studies in the series on the same platform. Data from 6 further submissions (2 for experiment 1A, 4 for experiment 2B) were removed since the email addressed supplied by the participant had been entered for a previous study. For experiment 1A, 40 minutes was taken as a minimum time limit consistent with conscientious attention to the stimuli, in the light of several pre-tests of the experiment (the data from which was not included in the analysis below). This criterion was consistent with a value of 42 minutes, calculated from the mean minus standard deviation of the actual experimental data. This time cut-off was then consistently applied to data from experiments 1B, 2A, and 2B. As noted below, experiments 3A and 3B were shorter, as half of the stimuli for phases 1, 3, and 4 (relating to scenarios not analysed here) were removed. The minimum time cut-off for experiment 3A, subsequently also

applied to 3B and to 4, was 30 minutes, based on pre-testing and the value of mean minus standard deviation for the actual data (30.08 minutes).

For each experiment, participant numbers were as follows (completed submission numbers in parentheses): 1A, N=47 (53), 23 male, 24 female, mean age 40.5 years, range 23 - 70 years; 1B, N=57 (72), 29 male, 28 female, mean age 34.03 years, range 19 – 73 years; 2A, N=65 (79), 41 male, 24 female, mean age 36 years, range 21 - 62 years; 2B, N=64 (77), 30 male, 34 female, mean age 33.4 years, range 19 - 63 years; 3A, N=46 (54), 31 male, 15 female, mean age 35.9 years, range 21 - 61 years; 3B, N=42 (50), 27 male, 15 female, mean age 28.9 years, range 18 - 69 years; 4, N = 42 (50), 17 male, 25 female, mean age 34.45, range 18 - 67.

Chapter 11 Replication results

11.1 A summary of the data on the response mode discrepancy

11.1.1 Errors and response mode compatibility overview

The data and figures given in this section are descriptive, without multi-level or any other kind of regression model. They simply show the proportions of normative responses, and also the proportion of responses for which there is a response mode discrepancy.

However, the data do take account of the question which arose in the previous section - for the delta mode, what response counts as neither discounting nor augmenting. That is to say, for which judgements in the delta mode is there no discrepancy with the response 'equally likely' in the change mode. In the model graphs in the previous section, 'equally likely' corresponded to a value of 'zero' for the predicted change rating. However, in delimiting the ROPE, discounting and augmenting were not assumed to be the present for every judgement other than 'zero'. This is intuitively more clear for the experiments from 2A to 4, where participants were asked to respond by indicating a position on a scale (a horizontal line) corresponding to a probability. As explained above, the indicator position was originally set in the middle of the line, and some interaction with the control was necessary for a participant to be allowed to move on with the experiment. A participant wishing to indicate the same value for the first response (R1) and the second (R2) would need some skill to actually get identical values, and a delta rating of 0. Thus, delta ratings in a range of +/1 0.1 were taken as 'practically equivalent' to zero, corresponding to the response 'equally likely' in the change mode. Discounting is, by this measure, a response from -1.0 to -0.2, while augmenting is a delta value from 0.2 to 1.0. The setting of an analogous range in experiments 1A and 1B was not so intuitively reasonable, since participants in the delta mode in those experiments could remember (presumably) their integer response for R1 and repeat it when making R2. However,

the assumption, right or wrong, is that in the delta mode participants did not so, but rather approached each judgement afresh. Thus, it was deemed preferable to treat experiments 1A and 1B analogously to the other experiments, and take -1, 0, and 1 delta scores as equivalent to the response 'equally likely' in the change mode. Discounting is a response from -10 to -2, while augmenting is a delta value from 2 to 10.

The proportions of normative (correct) answers for the two response modes and all seven experiments are shown in figures 6 and 7. The proportions of responses which matched in normativity between the two modes are shown in figure 8. Graphs for delta errors, change errors, and response compatibility for each scenario (with data combined from all seven experiments) can be found in Appendix 2.

Figure 6

Normative response proportions, delta, experiments 1A to 4

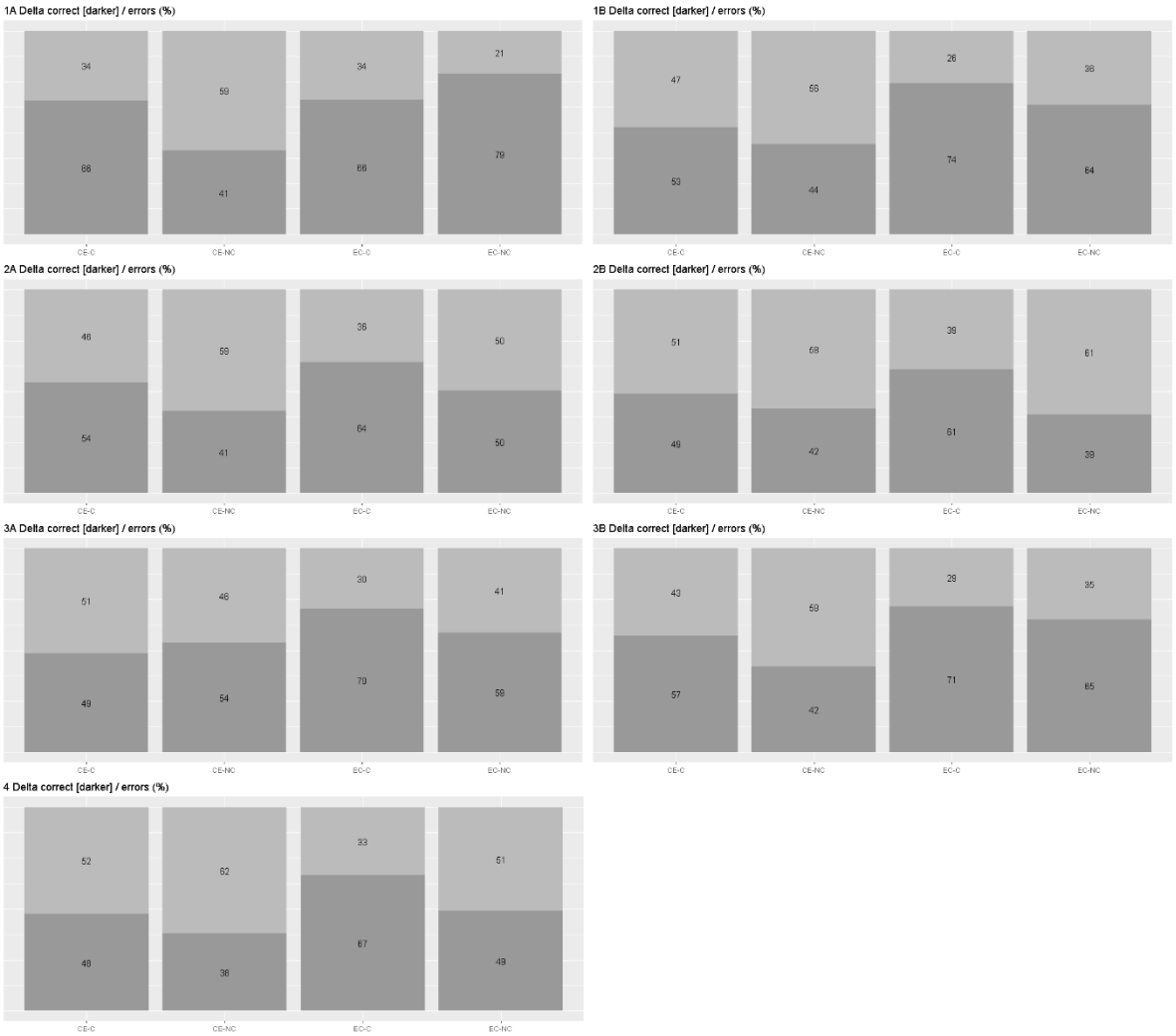
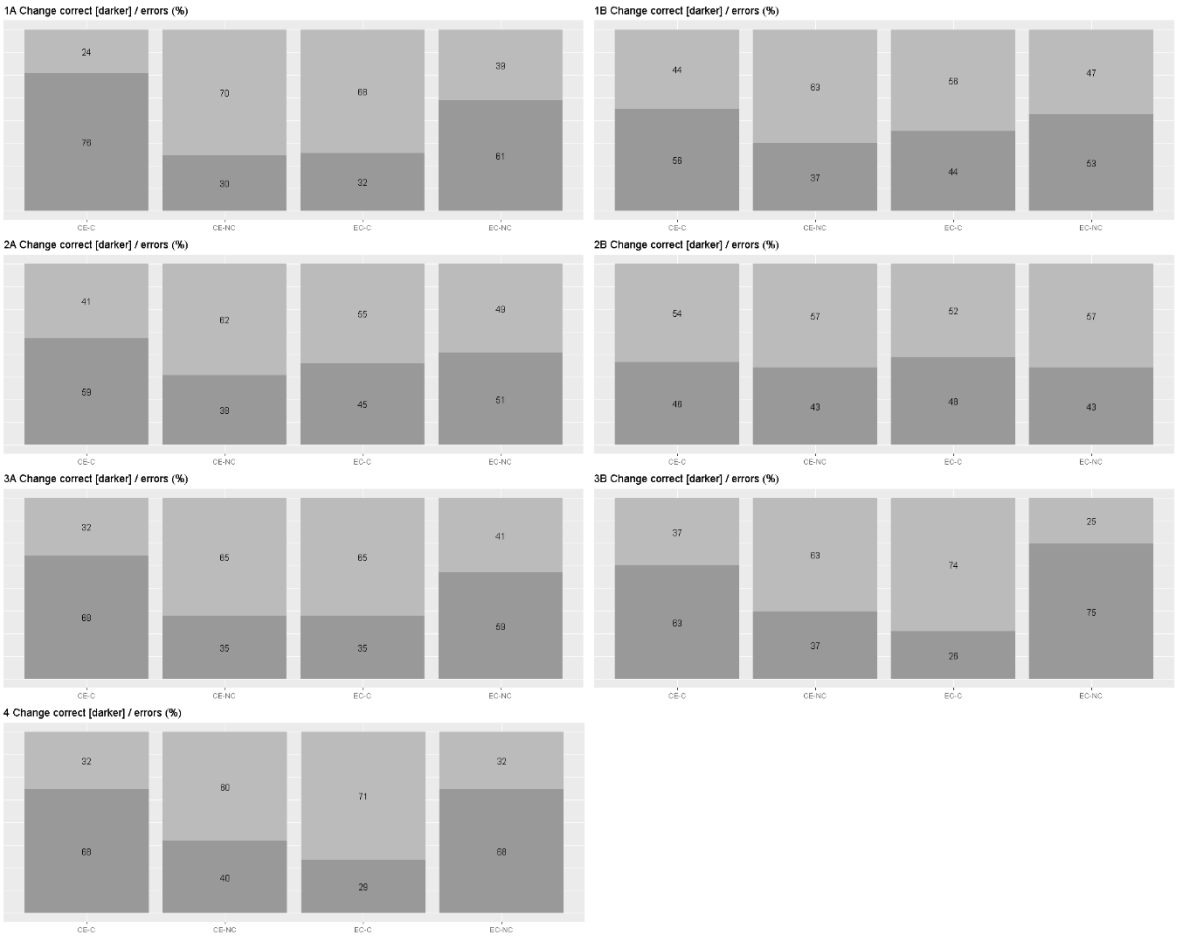


Figure 7

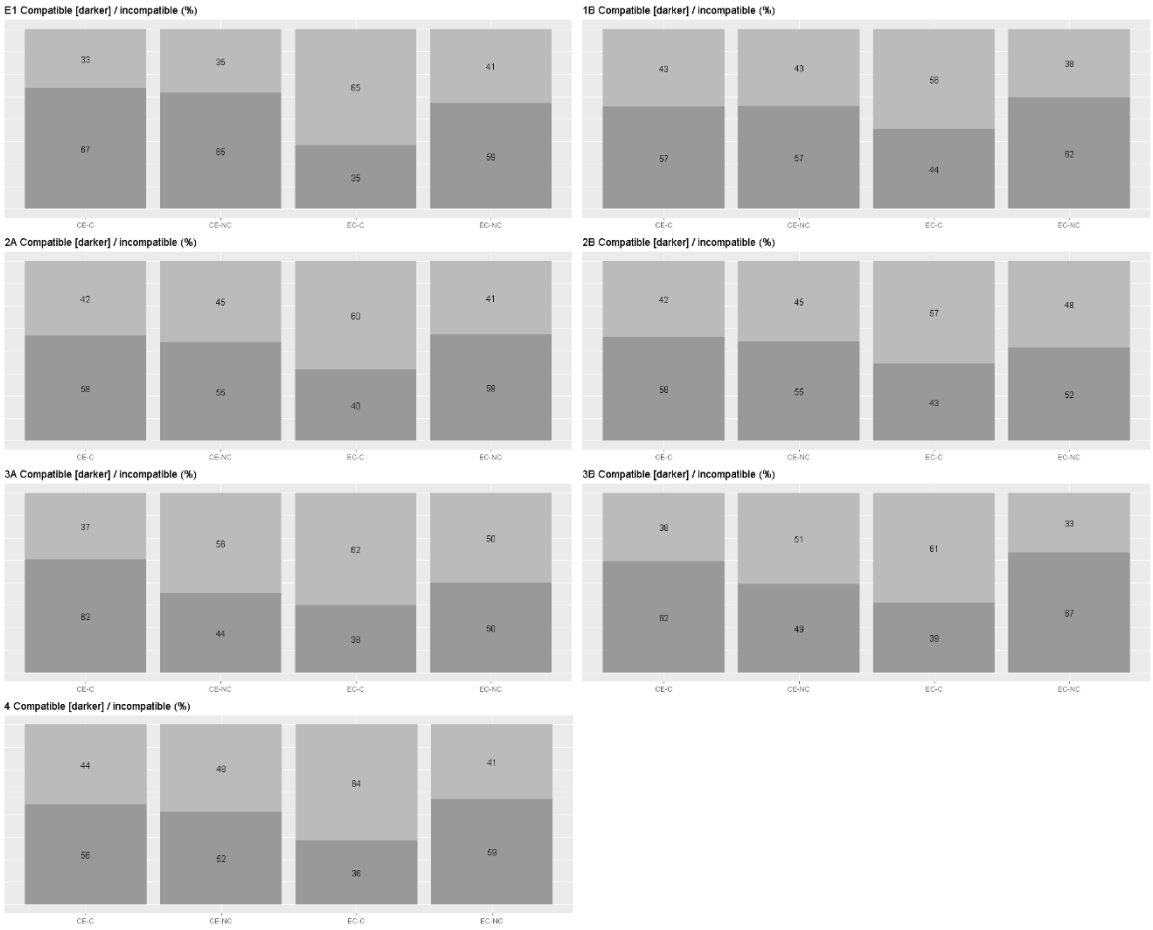
Normative response proportions, change, experiments 1A to 4



11.1.2 Errors and response mode compatibility by experiment

Figure 8

Compatible response proportions, experiments 1A to 4



11.1.2 Errors and response mode compatibility by experiment

Experiment 1A

For the change mode, participants gave correct answers in 50% of cases: 76% correct for CE-C, 30% for CE-NC, 32% for EC-C, and 61% for EC-NC. Overall, and for three of the four conditions, there were more correct responses for the delta response mode: 63% overall, 66% CE-C, 41% CE-NC, 66% EC-C, and 79% EC-NC.

Participants' judgements were more often 'compatible', that is to say, discounting, augmenting, or neither, for both change and delta, regardless of the normative response, for 57% of responses overall, and for three of the four conditions: 67% for CE-C, 65% for CE-NC, and 59% for EC-NC. The exception was EC-C, where only 35% of responses were compatible: this is the only condition in which there more 'correct' judgements in one response mode (delta), while in the other response mode (change) there were more errors. These results are shown in figure 8.

A chi-square test of independence showed the relation between experimental condition and compatibly normative responses was significant, χ^2 (3, N= 140) = 35.87, p<.001.

Experiment 1B

For the change mode, participants gave correct answers in 48% of cases: 58% correct for CE-C, 37% for CE-NC, 44% for EC-C, and 53% for EC-NC. Overall, and for three of the four conditions, there were more correct responses for the delta response mode: 59% overall, 53% CE-C, 44% CE-NC, 74% EC-C, and 64% EC-NC.

Participants' judgements were more often 'compatible' for 55% of cases overall, and for three of the four conditions: 57% for CE-C, 58% for CE-NC, and 62% for EC-NC. The exception was, as for experiment 1A, EC-C, where only 44% of responses were compatible: as for experiment 1A, this is the only condition in which there more 'correct' judgements in one response mode (delta), while in the other response mode (change) there were more errors.

A chi-square test of independence showed the relation between experimental condition and compatibly normative responses was significant, χ^2 (3, N= 171) = 11.65, p<.01.

Experiment 2A

For the change mode, participants gave correct answers in 49% of cases: 59% correct for CE-C, 39% for CE-NC, 45% for EC-C, and 51% for EC-NC. Overall, and for two of the

four conditions, there were more correct responses for the delta response mode: 53% overall, 55% CE-C, 41% CE-NC, 65% EC-C, and 51% EC-NC.

Participants' judgements were more often 'compatible', that is to say, discounting, augmenting, or neither, for both change and delta, regardless of the normative response, for three of the four conditions: 58% for CE-C, 55% for CE-NC, and 59% for EC-NC. The exception, as for experiments 1A and 1B, was once again EC-C, where only 39% of responses were compatible: this is the only condition in which there more 'correct' judgements in one response mode (delta), while in the other response mode (change) there were more errors.

A chi-square test of independence showed the relation between experimental condition and compatibly normative responses was significant, χ^2 (3, N= 195) = 19.2, p<.001.

Experiment 2B

For the change mode, participants gave correct answers in 45% of cases: 46% correct for CE-C, 43% for CE-NC, 48% for EC-C, and 43% for EC-NC. Overall, and for three of the four conditions, there were more correct responses for the delta response mode: 48% overall, 49% CE-C, 42% CE-NC, 61% EC-C, and 39% EC-NC.

Participants' judgements were more often 'compatible' for three of the four conditions: 58% for CE-C, 55% for CE-NC, and 52% for EC-NC. The exception was, as for experiments 1A, 1B, and 2A, EC-C, where only 43% of responses were compatible: as for experiment 1A, this is the only condition in which there more 'correct' judgements in one response mode (delta), while in the other response mode (change) there were more errors.

A chi-square test of independence showed the relation between experimental condition and compatibly normative responses was significant, χ^2 (3, N= 192) = 10.03, p=0.018.

Experiment 3A

For the change mode, participants gave correct answers in 49% of cases: 68% correct for CE-C, 35% for CE-NC, 35% for EC-C, and 59% for EC-NC. Overall, and for two of the

four conditions, there were more correct responses for the delta response mode: 58% overall, 49% CE-C, 54% CE-NC, 70% EC-C, and 59% EC-NC.

Participants' judgements were more often 'compatible', that is to say, discounting, augmenting, or neither, for both change and delta, regardless of the normative response, for two of the four conditions: 63% for CE-C, 44% for CE-NC, and equally compatible (50%) for EC-NC. The only condition, as for experiments 1A and 1B, 2A and 2B, with a majority of incompatible responses was EC-C, where only 38% of responses were compatible: this is the only condition in which there more 'correct' judgements in one response mode (delta), while in the other response mode (change) there were more errors.

A chi-square test of independence showed the relation between experimental condition and compatibly normative responses was significant, χ^2 (3, N= 138) = 19.28, p<.001.

Experiment 3B

For the change mode, participants gave correct answers in 50% of cases: 62% correct for CE-C, 37% for CE-NC, 26% for EC-C, and 75% for EC-NC. Overall, and for three of the four conditions, there were more correct responses for the delta response mode: 59% overall, 57% CE-C, 42% CE-NC, 71% EC-C, and 65% EC-NC.

Participants' judgements were 62% 'compatible' for CE-C, 49% for CE-NC, 39% for EC-C, and 67% for EC-NC.

A chi-square test of independence showed the relation between experimental condition and compatibly normative responses was significant, χ^2 (3, N= 126) = 24.06, p<.001.

Experiment 4

For the change mode, participants gave correct answers in 51% of cases: 68% correct for CE-C, 40% for CE-NC, 29% for EC-C, and 68% for EC-NC. Overall, and for three of the four conditions, there were more correct responses for the delta response mode: 48% overall, 48% CE-C, 38% CE-NC, 67% EC-C, and 38% EC-NC.

Participants' judgements were 56% 'compatible' for CE-C, 52% for CE-NC, 36% for EC-C, and 59% for EC-NC.

A chi-square test of independence showed the relation between experimental condition and compatibly normative responses was significant, χ^2 (3, N= 126) = 15.78, p=.001.

11.2 Replication results - Modelling discounting and augmenting

Multilevel regression models were fit predicting the delta and change ratings from the causal direction and whether the consequent was asserted.

11.2.1 Experimental order – model comparisons by experiment

Model comparisons were made using LOOICs (and WAICs for comparison) to show preferred models. The first comparisons were made between models with and without an effect of experimental order. The extra factor distinguished between the two experimental pathways in which phases 1 and 4 (see section 9.3 above) were delta and change, respectively, or vice-versa. These two orders were used merely as a counterbalancing precaution, and the model comparison was a check to see if order could safely be ignored in subsequent analyses. The model comparisons are summarised in table 2.

Using the notation commonly used to specify models for multi-level modelling software, the models compared here are:

1] A full model (with an effect of order)

*Change Rating / Delta Rating ~ Causal Direction*Assertion of Consequent*Experimental Order + (Causal Direction*Assertion of Consequent*Experimental Order | Participant) + (Assertion of Consequent*Experimental Order | Scenario)*

2] A model as above, but without an effect of experimental order

*Change Rating / Delta Rating ~ Causal Direction*Assertion of Consequent + (Causal Direction*Assertion of Consequent | Participant) + (Assertion of Consequent | Scenario)*

Note that in each case, the parentheses denoting the lower-level effects of scenario lacks a term for causal direction. This is because each scenario only exists for one causal direction (common effect or common cause). (Thus. in the models, presented below, with no consequent term, the scenario level effects are absent.

In these formulas, the value before the tilde is the dependent variable, and the values following the tilde are the independent variables. The vertical bar, |, separates fixed effects (before the bar) from random effects (after the bar). The plus sign adds effects without considering their interaction, while an asterisk shows effects which are modelled as interacting (Barr, et al., 2013).

Table 4

Comparisons for Models with and without an Effect of Experimental Order.

Model	Experiment	WAIC	WAIC weight	LOOIC	Bayesian stacking
Without an effect of order	1A (Delta)	2402.06	0.01	2411.41	0.29
With an effect of order	**1A (Delta)**	**2393.37**	**0.99**	**2402.53**	**0.71**
Without an effect of order	**1B (Delta)**	**3118.32**	**0.99**	**3124.51**	**1**
With an effect of order	1B (Delta)	3126.86	0.01	3134.45	0
Without an effect of order	**2A (Delta)**	**-265.78**	**0.99**	**-256.29**	**0.86**
With an effect of order	2A (Delta)	-255.48	0.01	-246.29	0.14
Without an effect of order	2B (Delta)	-63.47	0.06	-54.63	0.4
With an effect of order	**2B (Delta)**	**-69.13**	**0.94**	**-58.21**	**0.6**
Without an effect of order	**3A (Delta)**	**-59.17**	**0.88**	**-52.88**	**0.84**
With an effect of order	3A (Delta)	-55.20	0.12	-47.07	0.16
Without an effect of order	**3B (Delta)**	**-121.29**	**0.49**	**-117.95**	**0.73**
With an effect of order	3B (Delta)	-121.38	0.51	-114.82	0.27
Without an effect of order	**4 (Delta)**	**-68.51**	**0.99**	**-66.08**	**0.95**
With an effect of order	4 (Delta)	-60.13	0.01	-56.31	0.05
Without an effect of order	**1A (Change)**	**672.19**	**1**	**680.68**	**1**
With an effect of order	1A (Change)	683.54	0	697.03	0

Without an effect of order	**1B (Change)**	**1001.18**	**0.94**	**1007.38**	**1**
With an effect of order	1B (Change)	1006.84	0.06	1016.51	0
Without an effect of order	2A (Change)	1097.66	0.04	1105.50	0.28
With an effect of order	**2A (Change)**	**1091.33**	**0.96**	**1101.64**	**0.72**
Without an effect of order	**2B (Change)**	**1145.71**	**0.99**	**1152.32**	**1**
With an effect of order	2B (Change)	1155.59	0.01	1164.67	0
Without an effect of order	**3A (Change)**	**806.62**	**0.92**	**811.70**	**0.75**
With an effect of order	3A (Change)	809.12	0.08	816.54	0.25
Without an effect of order	**3B (Change)**	**763.01**	**0.91**	**769.42**	**0.86**
With an effect of order	3B (Change)	764.19	0.09	774.02	0.14
Without an effect of order	**4 (Change)**	**775.04**	**0.96**	**780.47**	**0.96**
With an effect of order	4 (Change)	779.46	0.04	786.97	0.04

Comparisons for models with and without an effect of experimental order. Preferred models in bold.

WAIC and LOOIC values for each model, along with the relevant model weights ('Bayesian stacking' for the LOOIC comparisons) are given in table 4. From the LOOIC values, for which a lower value indicates a preferred model, for all but 3 of the 14 experiments, the model without order was superior. The three exceptions were 1A and 2B in the delta mode, and 2A in the change mode. On the basis of these results, it was concluded that there was no reliable evidence that order had a significant effect, and the simpler model was used in all further analyses.

11.2.2 Model complexity – model comparisons by experiment

Comparisons were next made between three models of differing complexity: for each of the modes, firstly, the model with the rating as the outcome variable, predicted from causal direction (CE-EC) and consequent (C-NC) as fixed effects, and scenario and participant as random effects, and including an effect of the interaction of causal direction and consequent, secondly, the same model without the interaction effect, and thirdly, a model omitting consequent as a predictor. The results of these model comparisons are given in table 5.

The formulas specifying the three models for this comparison are:

1] A full model, as above (without an effect of order)

*Change Rating / Delta Rating ~ Causal Direction*Assertion of Consequent + (Causal Direction*Assertion of Consequent | Participant) + (Assertion of Consequent | Scenario)*

2] A model without interaction of causal direction and consequent

Change Rating / Delta Rating ~ Causal Direction+Assertion of Consequent + (Causal Direction+Assertion of Consequent | Participant) + (Assertion of Consequent | Scenario)

3] A model without an effect of consequent

*Change Rating / Delta Rating ~ Causal Direction + (Causal Direction*Assertion of Consequent | Participant)*

Table 5

Comparisons of models of varying complexity.

Model	Experiment	WAIC	WAIC weight	LOOIC	Bayesian stacking
Interaction, Consequent	**1A (Delta)**	**2402.06**	**0.99**	**2411.41**	**0.90**
No Interaction, Consequent	1A (Delta)	2411.26	0.01	2417.56	0.00
No Interaction, No Consequent	1A (Delta)	2514.68	0.00	2516.95	0.10
Interaction, Consequent	**1B (Delta)**	**3118.32**	**1.00**	**3124.51**	**0.89**
No Interaction, Consequent	1B (Delta)	3131.93	0.00	3136.70	0.00
No Interaction, No Consequent	1B (Delta)	3199.94	0.00	3201.40	0.11
Interaction, Consequent	**2A (Delta)**	**-268.20**	**1.00**	**-256.29**	**0.84**
No Interaction, Consequent	2A (Delta)	-239.69	0.00	-233.55	0.07
No Interaction, No Consequent	2A (Delta)	-156.61	0.00	-154.30	0.09
Interaction, Consequent	2B (Delta)	-63.47	0.18	-54.63	0.29
No Interaction, Consequent	**2B (Delta)**	**-66.45**	**0.82**	**-59.02**	**0.59**
No Interaction, No Consequent	2B (Delta)	-17.87	0.00	-15.23	0.12
Interaction, Consequent	**3A (Delta)**	**-59.17**	**1.00**	**-52.88**	**0.81**
No Interaction, Consequent	3A (Delta)	-41.74	0.00	-36.57	0.10
No Interaction, No Consequent	3A (Delta)	36.43	0.00	37.90	0.09
Interaction, Consequent	**3B (Delta)**	**-121.29**	**0.69**	**-117.95**	**0.56**

No Interaction, Consequent	3B (Delta)	-119.66	0.31	-117.55	0.40
No Interaction, No Consequent	3B (Delta)	-55.35	0.00	-54.53	0.04
Interaction, Consequent	4 (Delta)	-69.57	0.25	-66.64	0.52
No Interaction, Consequent	**4 (Delta)**	**71.78**	**0.75**	**-69.49**	**0.00**
No Interaction, No Consequent	4 (Delta)	-55.35	0.00	-54.53	0.48
Interaction, Consequent	**1A (Change)**	**672.19**	**0.92**	**680.68**	**0.66**
No Interaction, Consequent	1A (Change)	677.24	0.07	684.47	0.00
No Interaction, No Consequent	1A (Change)	683.91	0.00	688.55	0.34
Interaction, Consequent	1B (Change)	1001.18	0.24	1007.38	0.00
No Interaction, Consequent	**1B (Change)**	**998.82**	**0.76**	**1004.04**	**0.79**
No Interaction, No Consequent	1B (Change)	1013.27	0.00	1017.09	0.21
Interaction, Consequent	2A (Change)	1097.66	0.24	1105.50	0.00
No Interaction, Consequent	**2A (Change)**	**1095.40**	**0.74**	**1102.11**	**0.64**
No Interaction, No Consequent	2A (Change)	1102.00	0.03	1106.30	0.36
Interaction, Consequent	2B (Change)	1145.71	0.31	1152.32	0.18
No Interaction, Consequent	**2B (Change)**	**1144.69**	**0.51**	**1150.59**	**0.35**
No Interaction, No Consequent	2B (Change)	1146.79	0.18	1151.35	0.47
Interaction, Consequent	3A (Change)	806.62	0.38	811.70	0.22
No Interaction, Consequent	**3A (Change)**	**805.66**	**0.62**	**809.82**	**0.57**
No Interaction, No Consequent	3A (Change)	833.15	0.00	835.07	0.21

Interaction, Consequent	3B (Change)	763.01	0.11	769.42	0.00
No Interaction, Consequent	**3B (Change)**	**758.87**	**0.89**	**762.95**	**0.91**
No Interaction, No Consequent	3B (Change)	793.73	0.00	794.68	0.09
Interaction, Consequent	4 (Change)	775.87	0.42	781.35	0.53
No Interaction, Consequent	**4 (Change)**	**775.24**	**0.58**	**779.48**	**0.00**
No Interaction, No Consequent	4 (Change)	793.73	0.00	794.68	0.47

Model comparisons. Predictors: causal direction and consequent. Predicted Delta or Change ratings. Preferred models in bold.

When predicting the delta ratings, model comparisons preferred the fullest model, including the interaction between causal direction for five out of seven experiments. When predicting the change ratings, model comparisons preferred the simpler model without an interaction term, but retaining consequent as a fixed effect in six out of seven experiments. Figure 9 shows the results of the preferred models for each experiment for the delta mode, that is to say, including an interaction term. In these graphs experiments 1A and 1B have a scale reflecting the fact that for these experiments, participants gave values for each judgement (R1 and R2) ranging from 0 to 10.

Figure 10 shows the results of the preferred models for the change mode. The models gave probabilities of the three responses, ‘less’, ‘equally’, and ‘more’, from which expected values were calculated, taking each response to have a value of -1, 0, and 1, respectively.

Figure 9

The models (with an interaction term) for each experiment for the delta mode

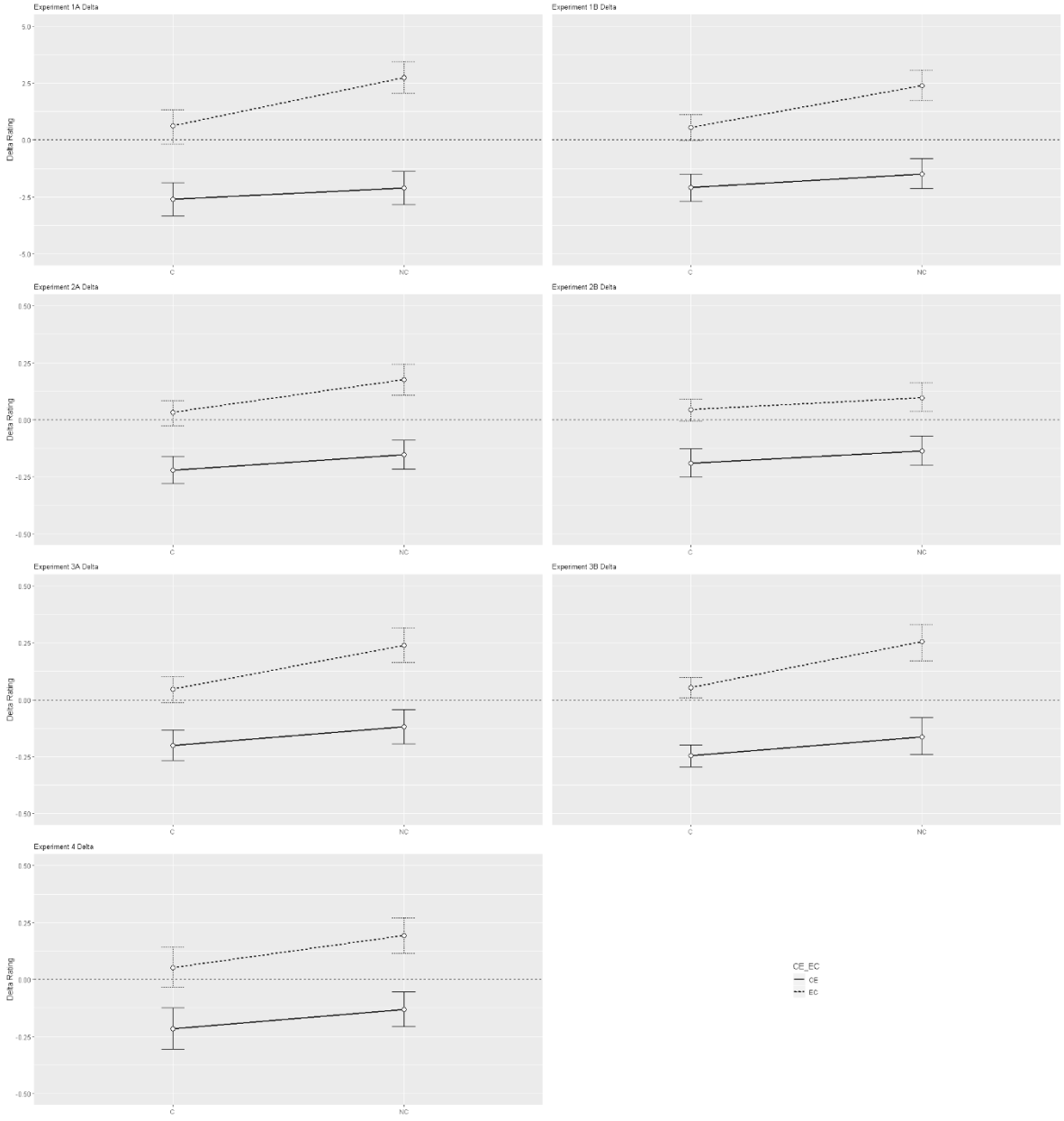
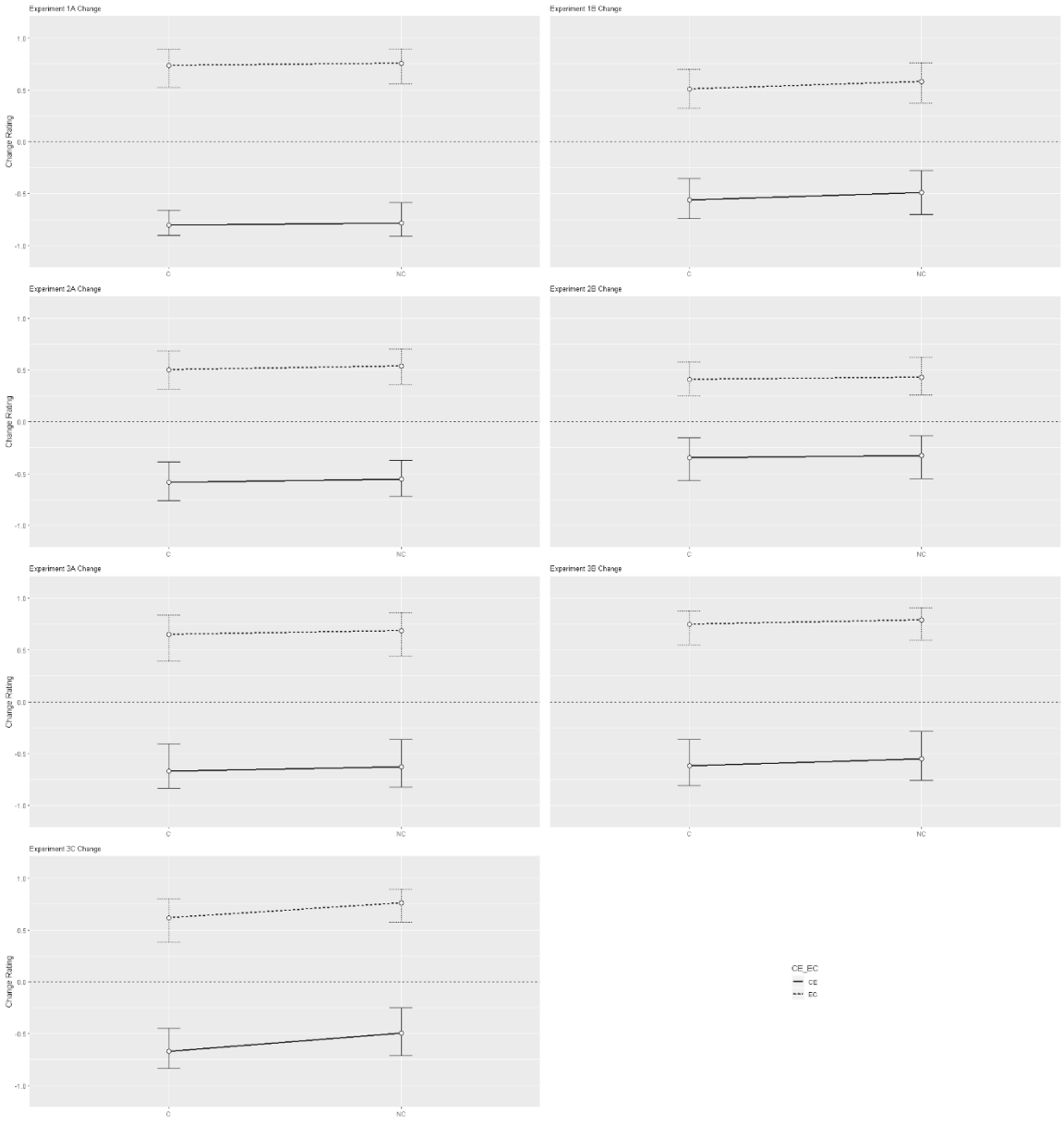


Figure 10

The models (without an interaction term) for each experiment for the change mode



11.2.3 Assessing effects with ROPE and HDI by experiment

The normative predictions for the 4 conditions (2 causal directions for consequent present or absent) are of discounting for the common effect structure when the consequent is asserted, and augmenting for the common cause structure when the consequent is not asserted. For the other two conditions, neither discounting nor augmenting is normative; in other words, the prediction here is of a null effect. While a null hypothesis cannot be statistically confirmed from a frequentist perspective, Bayesian methods can do so. To decide in each case whether the appropriate effect was observed, a region of practical equivalence (ROPE) and a highest density interval (HDI) of 95% (Kruschke, 2014) were used. The ROPE corresponding to a null effect (predicted for the EC-C and CE-NC conditions) was set to an interval of - 0.1 to + 0.1. When comparing the judgements given between the two response modes (see below, 'Errors and response mode compatibility') a similar criterion was used, taking values from -1 to +1 (experiments 1A and 1B) and - 0.1 to + 0.1 (experiments 2A to 4) in the delta mode to correspond to a rating of 'equally probable' in the change mode. For the comparisons here, where the change models give values as probabilities from - 1 to + 1, the same interval was used to correspond to 'no change'.

This choice of ROPE also addressed the fact that, for 5 of the 7 experiments, in the delta mode, response was by means of a slider control to which response was forced - with the control presented to participants centered on the scale, they were unable to move forward without interacting with the control. Although it was possible to do so by clicking exactly on the slider control, without moving the control to left or right, doing so required some deftness. In practice, an attempt to do so meant the slider was likely to move away from the centre, or no response was registered, requiring a further attempt before moving on. The most reliable way to give a response that did not move the control from the centre position was to move the slider away, and then re-centre it, an operation requiring observation of the numerical score feedback of the

slider position given at the far right of the scale. It seemed likely that in some cases participants wishing to respond with a value in the centre of the scale would make, deliberately or not, a small movement of the slider.

Following the model comparison carried out above, for the change mode, the models examined did not include an interaction term. For the delta mode, the full models, that is to say, including an interaction of consequent and causal direction, were examined.

For the change mode, across all seven experiments, and for all four experimental conditions, none of the 95% HDI fell within the ROPE. Thus, by this metric, the CE-C and EC-NC change mode responses were normative in all experiments, since they were not equivalent to zero, and the CE-NC and EC-C change mode responses were non-normative in all experiments, since zero change is normative for these conditions.

For the delta mode, in the CE-C condition, for all seven experiments, the HDI fell completely outside the ROPE, as is normative for this condition where discounting is predicted. For the condition where augmenting was expected, EC-NC, the HDI fell completely outside the rope for six experiments: 1A, 1B, 2A, 3A, 3B, and 4. For experiment 2B, 53.25% of the HDI fell within the ROPE.

For the two conditions where neither discounting or augmenting was normative, and for which the ROPE delimited normative responses, the results were mixed in the delta mode. For the CE-NC condition, the HDI fell completely outside the HDI for experiment 1A. For experiment 1B, 5.39% of the 95% HDI fell within the ROPE, for 2A, 2.26%, for 2B, 11.94%, for 3A, 29.97%, for 3B, 4.63%, and for 4, 16.71%. For the EC-C condition, the HDI fell completely inside the ROPE for experiments 2A, 2B, and 3B, in line with normative responses. For experiments 1A, 1B, 3A, and 4, most of the HDI fell within the rope: 1A: 86.64%, 1B: 95.87%, 3A: 99.24% and 4: 88.92%.

Effect sizes for the presence / absence of the consequent were estimated by calculating Cohen's d from the MCMC model samples for each experiment, response mode, and causal direction. For the common effect condition, effect sizes were larger for the delta response mode than for the change mode (delta / change, 1A: -1.32 [-1.37 -1.28]/ -0.31 [-0.35 -0.27], 1B: -1.89 [-1.95 -1.84]/ -0.65 [-0.69 -0.60], 2A -2.15 [-2.20 -2.09]/ -0.32 [-0.37 -0.28], 2B: -1.68 [-1.73 -1.63]/ -0.2 [-0.24 -0.15] 3A: -2.28 [-2.33 -2.22]/ -0.33 [-0.37 -0.29], 3B: -2.5 [-2.55 -2.44]/ -0.54 [-0.59 -0.5], 4: -2.04 [-2.09 -1.99]/ -1.55 [-1.60 -1.50]), and also for the common cause condition (delta / change, 1A: 5.78 [5.68 5.88]/ 0.25 [0.20 0.29], 1B: 5.89 [5.79 5.99]/ 0.67 [0.62 0.71], 2A: 4.62 [4.54 4.71]/ 0.35 [0.30 0.39], 2B: 1.92 [1.87 1.98]/ 0.24 [0.2 0.29], 3A: 5.51 [5.41 5.61]/0.3 [0.26 0.35], 3B: 6.25 [6.14 6.36]/ 0.5 [0.45 0.54], 4: 3.46 [3.39 3.53]/ 1.49 [1.45 1.54]).

11.2.4 Analysis of combined data set – model comparisons

After these analyses by experiment, the data from the experiments were combined, and models fit to this full data set.

The model comparisons for the discounting / augmenting analyses by experiment, with and without an effect of order, were considered sufficiently conclusive to make a repetition of such a comparison for the combined data set unnecessary.

The comparisons of the three models of varying complexity (with /without an interaction term, with / without an effect of assertion of the consequent) were repeated for this combined data set. The results of this model comparison are summarised in table 6.

Table 6

Model Comparisons for the Combined Data Set

Model	Experiment	WAIC	WAIC weight	LOOIC	Bayesian stacking
Interaction, Consequent	All (Delta)	-157.08	0.34	-156.30	0.19
No Interaction, Consequent	**All (Delta)**	**-158.44**	**0.66**	**-157.80**	**0.77**
No Interaction, No Consequent	All (Delta)	213.59	0.00	213.59	0.04
Interaction, Consequent	All (Change)	7279.33	0.35	7280.05	0.00
No Interaction, Consequent	**All Change)**	**7278.07**	**0.65**	**7278.72**	**0.95**
No Interaction, No Consequent	All (Change)	7399.92	0.00	7400.38	0.05

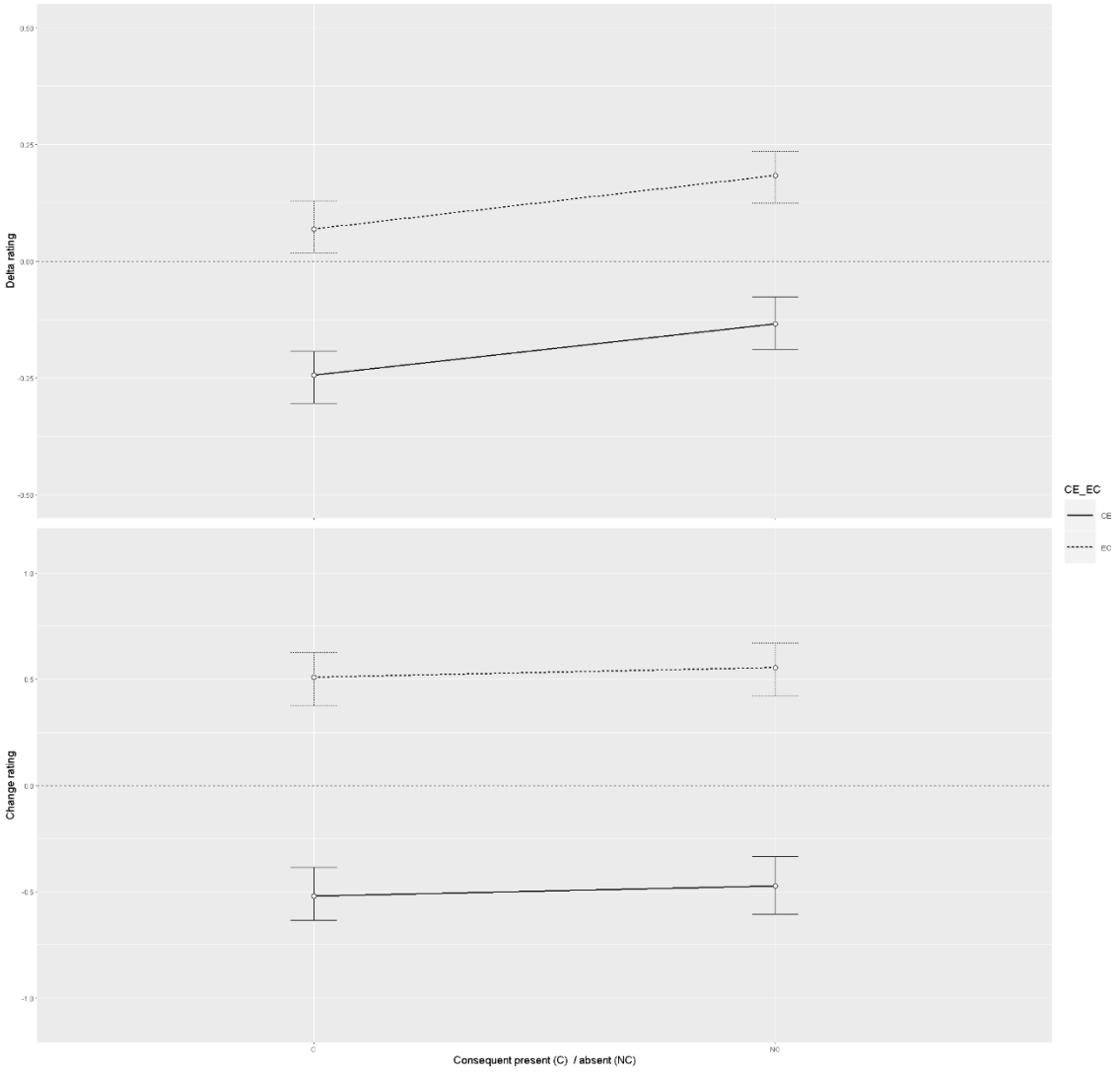
Model Comparisons for the Combined Data Set. Predictors: causal direction and consequent. Predicted Delta or Change ratings. Preferred models in bold.

The model formulae for these comparisons are as given above. When predicting both the delta and change ratings, model comparisons the simpler model without an interaction term, retaining consequent as a fixed effect. Thus the preferred model for the combined set, without an interaction term, was different from that chosen on the basis of the majority preference for the delta response mode (including the interaction term). The preferred model, that without an interaction term was the same for the change response mode as for a majority of the per-experiment comparisons.

Figure 11 shows the results of the model without interaction for each response mode.

Figure 11

The preferred models for the combined data set: delta (above), change (below) without an interaction of causal direction and with a term for the presence of the consequent



11.2.5 Analysis of combined data set – ROPE and HDI

As described above, a ROPE was set for each condition, and how much of the HDI fell within that ROPE was calculated. The differences from the analyses given above were as follows. Firstly, for both modes, the models examined were without an interaction term, in line

with the model comparison. Thus, the model for the delta mode was different from those used when examining each experiment separately. Secondly, the delta scores for the experiments 1A and 1B were adjusted to match the scale of the subsequent experiments while collating the individual data sets. This meant that a ROPE was chosen as described above for experiments 2, 3, and 4 in the delta mode.

For the change mode, for all four experimental conditions, none of the 95% HDI fell within the ROPE. Thus, as for the individual analyses, by this metric, the CE-C and EC-NC change mode responses were normative in all experiments, since they were not equivalent to zero, and the CE-NC and EC-C change mode responses were non-normative in all experiments, since zero change is normative for these conditions.

For the delta mode, in the CE-C condition, the HDI fell completely outside the ROPE, as is normative for this condition where discounting is predicted. For the EC-NC condition, the HDI also fell completely outside the rope. This is also normative for this condition, where augmenting is expected.

Once more, results were not conclusive for the two conditions where neither discounting or augmenting was normative, and for which the ROPE delimited normative responses. For the CE-NC condition, 0.09% of the 95% HDI fell within the ROPE. For the EC-C condition, 90.0% of the 95% HDI fell within the (normative) ROPE.

Effect sizes for the presence / absence of the consequent were again estimated by calculating Cohen's d from the MCMC model samples for each experiment, response mode, and causal direction. For the common effect condition, effect sizes were larger for the delta response mode than for the change mode (delta / change,: -3.98 [-4.06 -3.91] / -0.71 [-0.75 -0.66]), and also for the common cause condition (delta / change, : 4.01 [3.93 4.08] / 0.71 [0.67 0.76]).

11.2.6 Discussion of the models

For the change mode, the results are clear and unambiguous (see table 8). For every experiment, discounting and augmenting are found when they should be (CE-C and EC-NC). Discounting and augmenting are also found for every experiment in the conditions where they are inappropriate (CE-NC and EC-C).

For the delta mode, the results are mixed (see table 7). As for the change mode, when discounting and augmenting are normative, they were found. There is one exception – in experiment 2B, in the EC-NC condition, the result is not clear, with 53% of the ROPE falling within the HDI. For EC-C, normative behaviour was found for only 3 of the 7 experiments. However, it is worth looking at the delta results more closely.

Table 7

Delta Mode Results

	Normative	1A	1B	2A	2B	3A	3B	4	All 7
CE-C	Disc	Disc	Disc	Disc	Disc	Disc	Disc	Disc	Disc
CE-NC	Neither	Disc	Und	Und	Und	Unc	Und	Und	Und
EC-C	Neither	Und	Und	Neither	Neither	Und	Neither	Und	Und
EC-NC	Aug	Aug	Aug	Aug	Und	Aug	Aug	Aug	Aug

Delta mode results. Discounting (‘Disc’), Augmenting (‘Aug’), Neither. Undecided (‘Unc’) = 95% HDI partially within a ROPE. Normative results shaded grey.

Table 8

Change Mode Results

	Normative	1A	1B	2A	2B	3A	3B	4	All 7
CE-C	Disc	Disc	Disc	Disc	Disc	Disc	Disc	Disc	Disc
CE-NC	Neither	Disc	Disc	Disc	Disc	Disc	Disc	Disc	Disc
EC-C	Neither	Aug	Aug	Aug	Aug	Aug	Aug	Aug	Aug
EC-NC	Aug	Aug	Aug	Aug	Aug	Aug	Aug	Aug	Aug

Change mode results. Discounting ('Disc'), Augmenting ('Aug'), Neither ('None').

Undecided ('Unc') = 95% HDI partially within a ROPE. Normative results shaded grey.

The rationale behind the use of a ROPE is to be able, when all or none of it falls within a chosen HDI, to say with confidence that an effect is, or is not, supported by the data. However, used in this way, the procedure can be seen as re-introducing, via the back-door, the logic behind NHST statistics. Although Bayesian procedures have a better theoretical basis when used for examining hypotheses on the basis of data, the judgements made in the process of reaching a yes/no decision (as seen in the use of the term 'equivalence test' for this procedure, with the associated terms 'accepted', 'rejected', and 'undecided'. The percentage of the HDI most often recommended for use with a ROPE is 89% (Kruschke, 2014; Makowski, Ben-Shachar, Lüdecke, 2019; McElreath, 2015). There is said to be a danger that MCMC calculations with a 95% HDI will be at risk of sampling artefacts unless more than 10,000 samples are drawn (in MCMC terminology, the calculation will be 'unstable', Kruschke, 2014). Since the results reported here are intended as a replication of earlier research using 95% confidence intervals (Ali et al., 2010; Ali et al., 2011) or credible intervals (Hall et al., 2016) a

95% HDI is reported here. The equivalent values for an 89% HDI, which differ little, are reported in appendix 3 for comparison. The ROPE itself was chosen as described above, so as to interpret as equivalent to 'no change' values that were not exactly zero for the delta mode. Interpreting the values of the analyses presented here requires, to take advantage of the richer information provided by Bayesian analyses, in contrast to traditional methods, looking at how much of the ROPE falls within the HDI for the values presented in tables 5 and 6 as 'undecided'.

For the delta mode, the EC-C condition is normative for 3 out of 7 experiments. However, for the other 4 experiments, for which the result 'undecided' is given by the ROPE test, virtually all (96%, experiment 1B, and 99%, experiment 3A) or most (87%, experiment 1A, and 88%, experiment 4) of the ROPE fell within the HDI delimiting normative judgements. For the CE-NC condition, 1A is non-normative, and for the others, the results, all 'undecided', can be said, using NHST terminology, to 'approach' non-normativity. For only one of the experiments, 3A, does the percentage of the ROPE within the HDI exceed 20%. For these two conditions in the delta response mode where participants' judgements are not clearly normative or non-normative, density plots are shown in figure 12 for CE-NC, and figure 13 for EC-C (data from experiments 1A and 1B were put on the same scale as that from the other experiments in plotting these graphs).

These data can fairly be interpreted, as suggesting some discounting for CE-NC, and mostly normative behaviour, with very minor augmenting, for EC-C.

Figure 12

Density plots for the delta mode, condition CE-NC

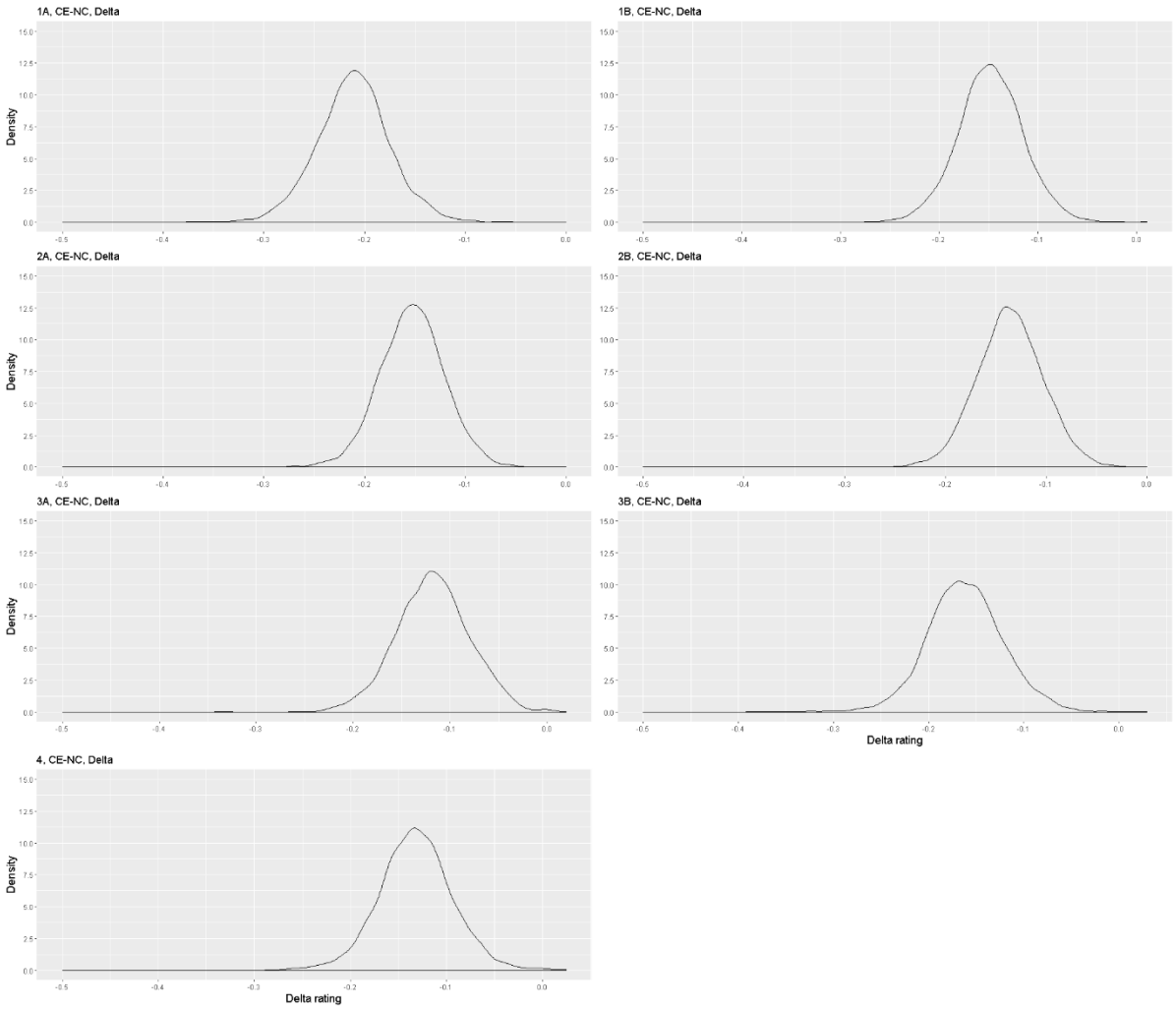
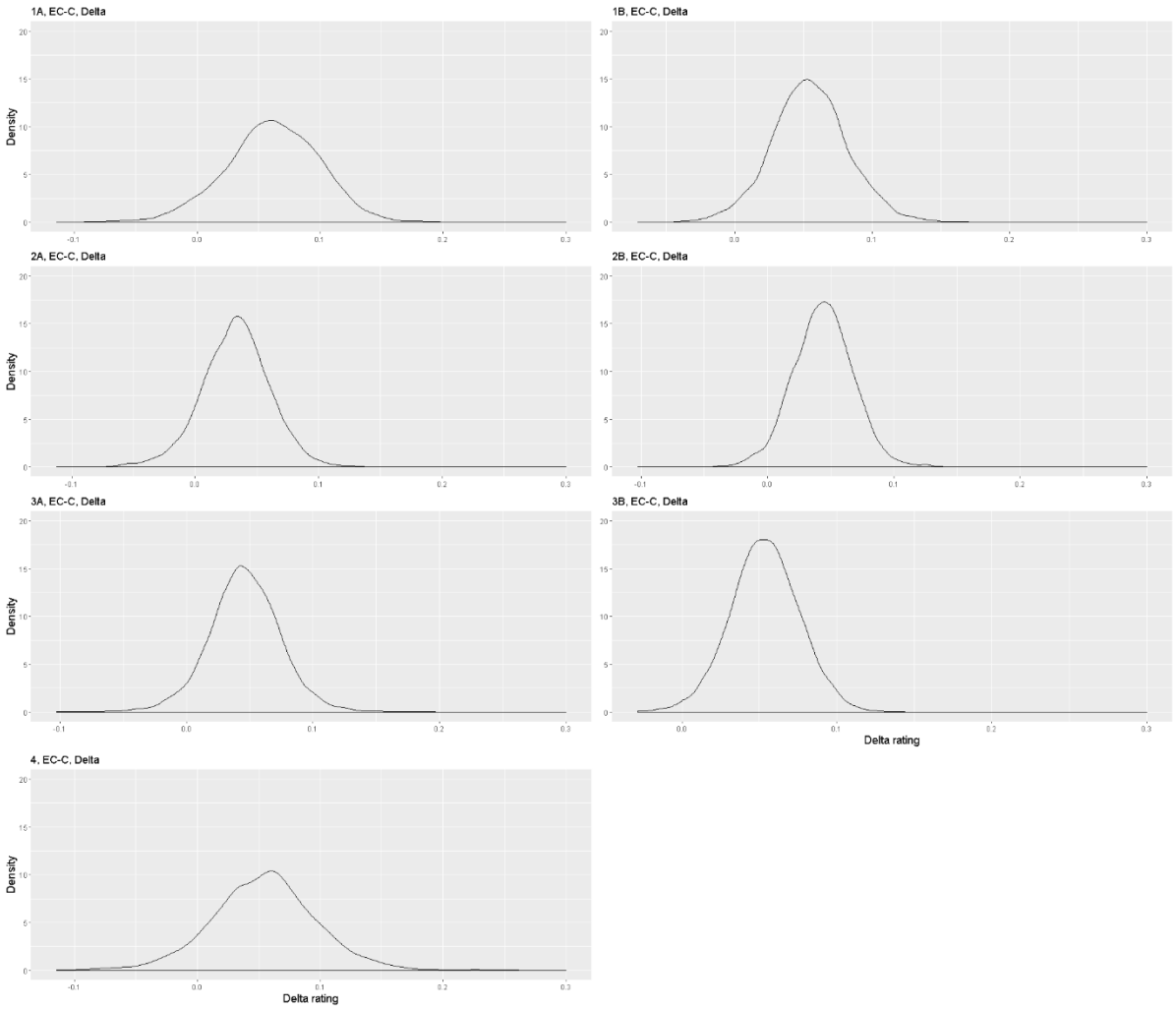


Figure 13

Density plots for the delta mode, condition EC-C



To sum up the results of these replications, taking the seven experiments separately, the results for the change mode are a clear replication of Ali et al. (2011), and Hall et al., (2016). Discounting was found regardless of the consequent for common effect scenarios, and augmenting, regardless of the consequent, for common cause scenarios.

For the delta mode, for common effect scenarios, discounting was found in every experiment for which it was appropriate, that is to say, when the consequent (the effect) was asserted. For common cause experiments, with the exception of one experiment, augmenting was found when appropriate, i.e., when the consequent (the cause) was not asserted.

For the delta mode, for the common cause scenarios in condition EC-C, several experiments showed clear evidence of neither augmenting or discounting, according to the slightly extended criteria (a range of -1 to 1 for experiments 1A and 1B, and -0.1 to 0.1 for the others). The remaining experiments showed a clear tendency for this non-normative behaviour. This is a replication of the earlier experiment, with child participants, Ali et al., (2010), and not of the subsequent Ali et al., (2011).

For the common effect scenarios, in condition CE-NC, the results were ambiguous. There is a tendency to discount, but the results do not allow the conclusion that the results of either Ali et al., (2010), where discounting was found, nor those of Ali et al., (2011), where normative behaviour was found, were replicated.

With the data from all experiments collated, the models showed similar results. All the change mode conditions, according to the ROPE based equivalence tests, replicated the earlier research. The same was the case for the CE-C and EC-NC conditions in the delta mode. For EC-C, the results suggested a replication of the (normative) results of Ali et al., (2010), but the effect just failed the ROPE equivalence test. For CE-NC, the participants' behaviour was not normative, but failed to show clear discounting-like behaviour.

11.3 Discussion

For the change mode, the models fit showed discounting for the CE-C and CE-NC conditions, and augmenting for the EC-NC and EC-C conditions. The results for CE-NC and EC-C are not normative (and should more precisely be called 'discounting-like' behaviour for CE-NC, and 'augmenting-like' behaviour for EC-C). The proportions of correct and incorrect responses reflect the models. For CE-C, every experiment except 2B showed more normative responses (i.e., most answers were 'less likely'). For EC-NC, experiment 2B was also the only one to not produce a majority of normative responses (i.e., 'more likely').

For the two other change conditions, CE-NC and EC-C, no experiment produced a majority of normative responses, in line with the fitted models.

For the change mode, it should be remembered that the models and the raw data shown in this section are not entirely compatible, inasmuch as here responses are only categorised as 'correct' and 'incorrect'. The categorical models fit to the change data produced response probabilities which were used to produce predicted responses on a scale from -1 to 1. In this sense, for the model fits, 'more likely' is a more non-normative response for the CE-C condition than 'equally likely'. For example, for CE-C, for experiment 1A, 5 out of 141 responses were 'extra' non-normative in this sense. For the other experiments the figures were: 1B 13/171, 2A 11/195, 2B 31/192, 3A 9/138, 3B 14/126, 4 12/126.

The delta mode data reveals a more complicated picture. As for change, the fit models show appropriate behaviour for CE-C (discounting) and EC-NC (augmenting). For CE-C, in 3 experiments, a majority of responses were not normative (taking into account the restricted definition of discounting and augmenting described above). However, in these three cases the normative responses were at or above 48%. For EC-NC, 5 experiments had a majority of normative responses, experiment 4 was at 49%, and 2B was again a striking outlier, with EC-NC receiving the lowest proportion of normative responses amongst the 4 conditions.

For the delta mode, the EC-C condition produced a majority of normative responses for every experiment. The model fits reported above showed strong, if not completely consistent, support for normative (neither discounted nor augmented) judgements in this condition. With one exception, experiment 3A, the CE-NC condition showed non-normative (either discounting or augmenting) behaviour, though not at levels as high as those for this condition or EC-C in the change mode.

The third measure reported here is judgement consistency – if participants gave normative responses in one mode, did they do so in the other? Not doing so indicates the

response mode discrepancy, and the model fits reported above indicate the crucial condition for this discrepancy is EC-C, with normative behaviour in the delta mode, but not in the change mode. Indeed, responses were mostly non-compatible for every experiment in this condition, never reaching a level of 45%, while in the other conditions, only experiments 3A and 3B in the CE-NC condition failed to show a majority of compatible responses.

The χ^2 tests, although not Bayesian, are reported above and show a significance difference in compatibility for the 4 conditions in every experiment.

Thus, the response mode discrepancy is supported both by the raw data, and by the multi-level model fits.

Chapter 12 Results which are not replications of earlier research

12.1 Working memory

12.1.1 WM scores as a fixed effect

Research has shown that individuals with higher WM capacity perform better on conditional reasoning tasks (Handley, Capon, Copp & Harper; Toms, Morris, Ward, 1993). MMT puts forward a specific explanation for this; individuals with lower WM capacity are less able to represent all the models associated with a particular reasoning task (Johnson-Laird, 1983). As described above, in phase 2 of each experiment, participants were asked to carry out a task as measure of their WM capability. The task consisted of 16 questions, giving participants a WM score ranging from 0, for those answering all questions incorrectly, to 16, for those who gave all correct answers.

Across the seven experiments, 363 participants produced 362 WM scores (the score for one participants in experiment 3A was not recorded by the Qualtrics platform). The scores were distributed bimodally (see figure 14), which is unexpected. It seems unlikely that the participants' actual WM capacities had a bimodal distribution. Figure 15 shows the distribution of scores which would result from participants guessing their answers. The mode of this distribution closely matches the lower of the two modes of the distribution of the experimental WM scores. Phase 2, which was also intended to provide a buffer between phases 1 and 2, reducing interference from the same scenario (though in a different condition of consequent) judged in phase 1 and again in phase 4, was in fact tedious and repetitious, and it seems likely that many of the participants failed to engage conscientiously, and merely answered at random. (Conversely, while participants were asked to use only their memory, and not make a note of the WM tasks as they were carried out, for example with a pen and paper, there was no way of checking that participants complied with the request.)

Fitting models to the data for each particular score required combining the datasets, as there were insufficient datapoints, particularly for the middle scores, for the sampling to converge. These models were fit in brms with causal direction and consequent as fixed effects, and scenario and participant as random effects (to achieve an acceptable effective sample size (ESS), it was necessary to sample from 4000, rather than 2000, iterations for the delta mode results). Figures 16 and 17 show plots for separate model fits for participants at each WM score. Ratings, and thus discounting and augmenting (whether appropriate or inappropriate) are higher for both response modes for higher WM scores. This does not appear to be a linear trend, but rather a higher level of judgements for the scores making up the higher of the two modes in the distribution of scores (figure 14). Possibly participants who did not answer the WM task conscientiously were also guessing, and hedging their bets, on the causal conditional tasks in phases 1 and 4.

Figure 14

Number of participants by WM score across all seven experiments

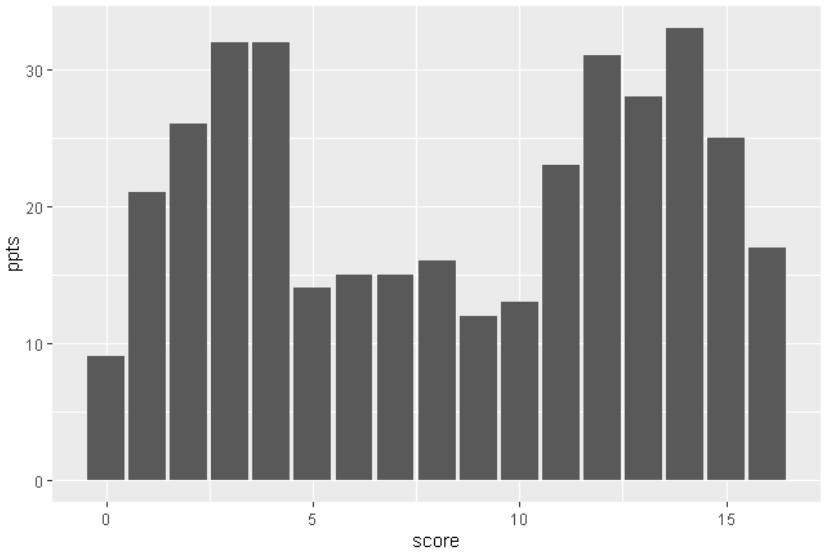


Figure 15

Probabilities of guessing correct answers in the WM task

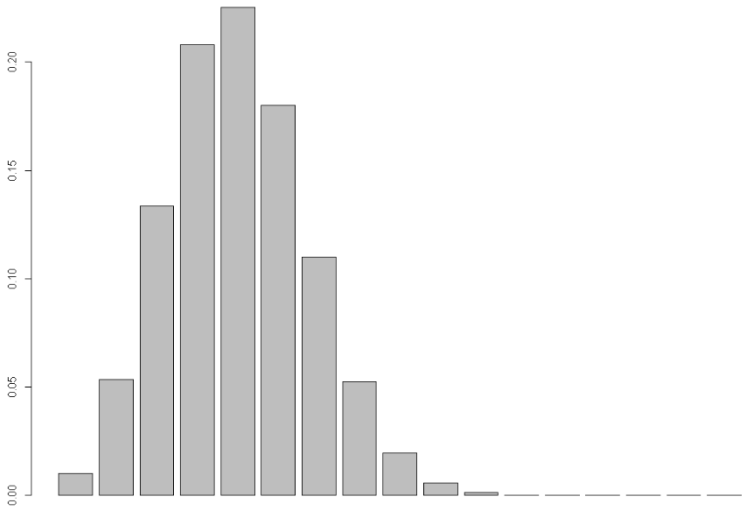


Figure 16

Individual models fit to the delta mode data for participants at each WM score (all seven experiments)

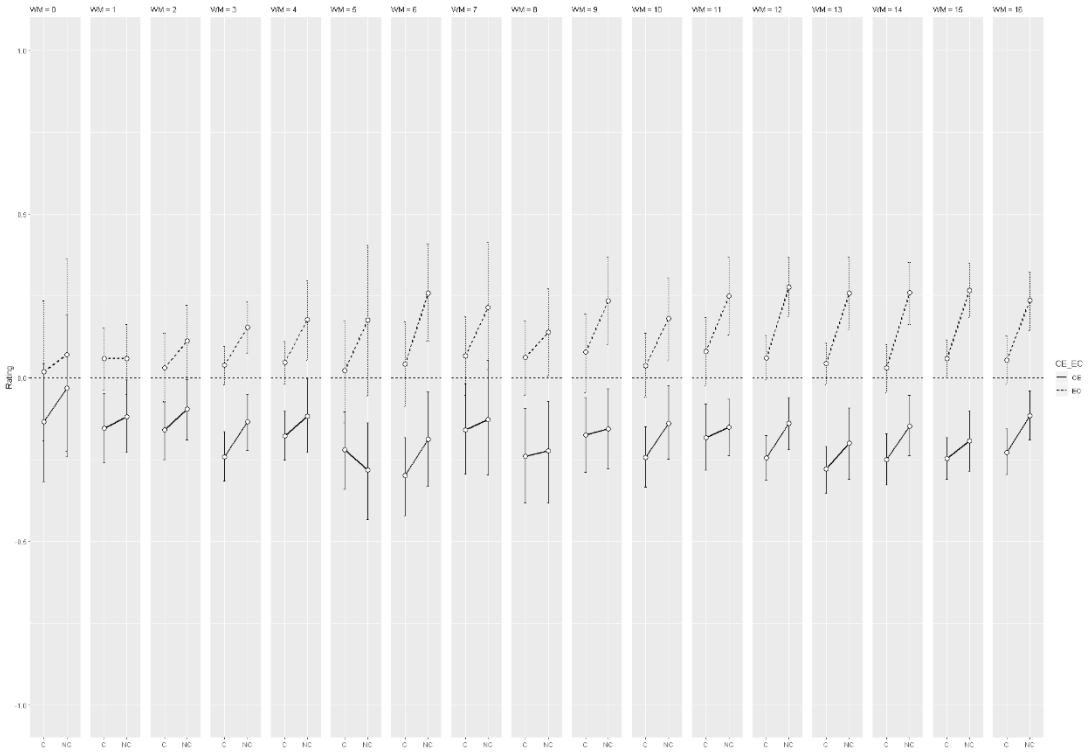
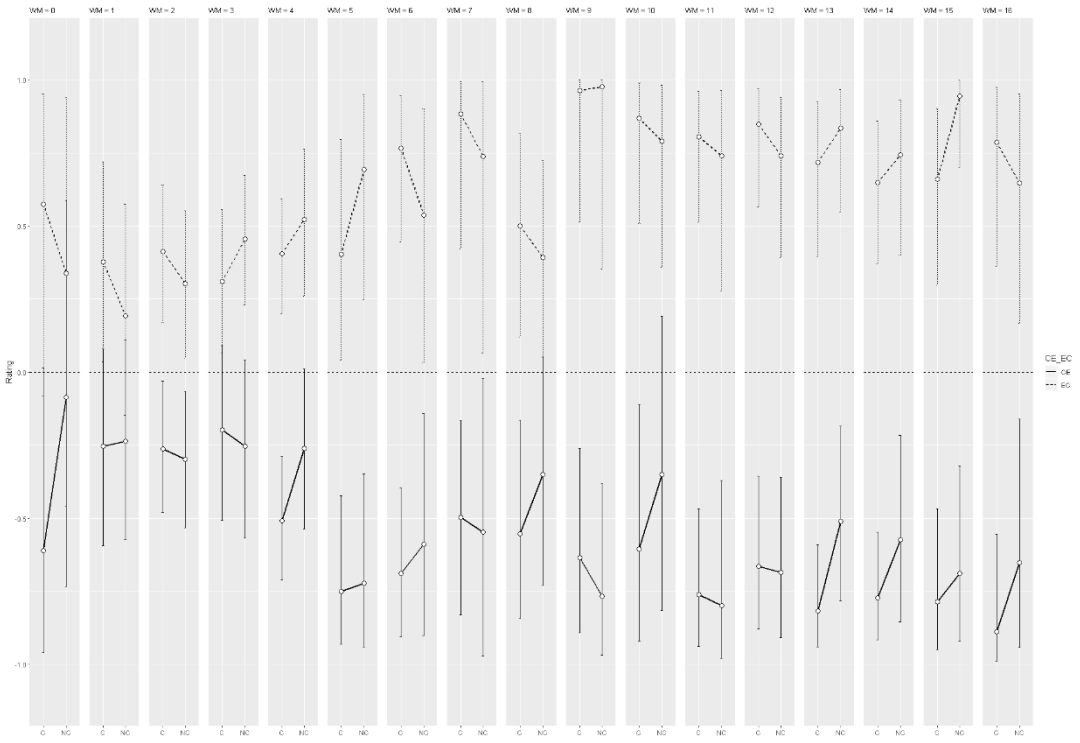


Figure 17

Individual models fit to the change mode data for participants at each WM score (all seven experiments)



Models were fit in brms to the judgements of participants at each of the 17 possible scores. These models were fit in brms with causal direction and consequent as fixed effects, and scenario and participant as random effects (to achieve an acceptable effective sample size (ESS), it was necessary to sample from 4000, rather than 2000, iterations for the delta mode results). Figures 16 and 17 show plots for separate model fits for participants at each WM score.

12.2 WM scores and a normativity metric

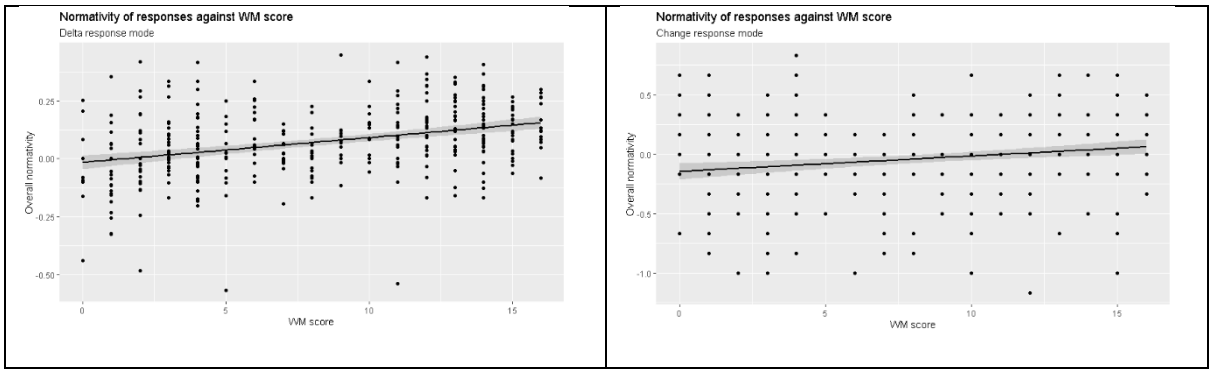
A ‘normativity metric’ was calculated to compare response normativity for the participants showing the different WM scores by subtracting the absolute value of discounting

/ augmenting in conditions for which neither was normative (CE-NC and EC-C) from discounting / augmenting in those conditions where one of those was normative (CE-C for discounting and EC-NC for augmenting), reversing the sign of the CE-C condition value, to match that of EC-C, so as to achieve a metric for both causal directions for which a higher score indicated greater overall normativity. Plots of this normativity metric against participants' WM scores for both response modes are shown in figure 18. Calculating this metric requires a value for both consequent conditions (asserted – C, and not asserted – NC), but only one judgement was obtained for each participant for each scenario (see experimental procedure above). Thus, the average of the three scenarios seen by each participant for each consequent condition was used for calculations. Although the change mode responses were ordinal (answers were 'less likely', 'equally likely', and 'more likely'), the metric of normativity used here is calculated arithmetically from those responses coded as -1, 0, and 1, and thus brms models were fitted for both change and delta modes taking normativity as a ratio variable.

For the delta mode, increasing WM scores were associated with greater normativity of judgement, and the correlation was significant ($r = 0.38$, $N = 358$, $p < 0.001$), and for the change mode, the correlation was also significant ($r = 0.09$, $N = 360$, $p = 0.0002$). (Different Ns are dues to failures by Qualtrics to record values.)

Figure 18

Normativity metric of models for participants at each WM score, delta and change modes



For each causal direction a model was fitted predicting values of the normativity metric from the WM score, and was compared it to a model predicting the metric from a dummy variable (1). In both cases, the model including WM score as a factor was shown to be preferable (see table 9).

Table 9

Model Comparisons Showing WM Score is a Predictor of Judgement Normativity

Model	Response mode	WAIC	WAIC weight	LOOIC	Bayesian stacking
Without an effect of WM score	Delta	-322.77	0	-322.77	0.07
With an effect of WM score	**Delta**	**-366.82**	**1**	**-366.83**	**0.93**
Without an effect of WM score	Change	239.62	0	239.62	0.23
With an effect of WM score	**Change**	**227.91**	**1**	**227.92**	**0.88**

Model comparisons showing WM score is a predictor of judgement normativity (see above for explanation of metric used). Preferred models shown in bold.

12.1.3 Discussion

These results do suggest that participants with a greater WM capacity were more likely to reason normatively. However, the striking and unexpected bimodal distribution of the scores suggests that the method of assessing WM capacity was flawed, in the sense that it led many participants not to complete the (long, repetitive, and tedious) task conscientiously. Although the length of the WM task (16 repetitions) was deliberate, and intended to provide a buffer between the two presentations (in phases 1 and 4) of the scenarios, it seems that a better choice might have been to assess WM with fewer questions, and provide other filler material.

12.2 Confidence and response time – evidence for conflicted reasoning

12.2.1 The rationale

Experiment 2 in Hall et al. (2016) considered a particular heuristic, related to MMT, which might lie behind participants' non-normative reasoning in relation to discounting and augmenting. Hall et al. suggested that reasoners might be constructing a simple model which ignored the subsequent presentation of the consequent. That is to say, for a common effect scenario, participants might simply assess the probability of exactly one cause (i.e., exclusive-OR), and for a common cause scenario, the probability of either, or both, effects (i.e., the conjunction). Such a representation corresponds to an initial model, subject to expansion / improvement given a lack of conviction in the model's accuracy, and given time and cognitive resources.

Such 'rough-and-ready' reasoning is a bias, or more positively, a heuristic. Evidence suggests that reasoners may not simply use a heuristic, but rather, may make a beginning with a fuller strategy in parallel. This may mean that they have some knowledge of when their reasoning is not optimal, as opposed to when the heuristic result is also the normative result.

De Neys (2012) reviewed evidence from response times and eye-tracking measures that, for problems where a heuristic strategy produced a 'biased' result, reasoners spent longer and attended differently than for problems where the heuristic gave a non-biased result. De Neys, Cromheeke, and Osman (2011), Gangemi, Bourgeois-Gironde, and Mancini (2015), Janssen, Velinga, de Neys, and van Gog (2021), and Mevel et al. (2015), provide evidence that reasoners have some conscious access to knowledge that the answer they have just given to a problem is somehow incorrect; in these studies the participants were asked to express their confidence in their judgements. Janssen et al. (2021), using scenarios longer than those in earlier research (and more in line with the complexity of the scenarios used in this research) found conflict detection effects for confidence judgements, but not for reaction times.

12.2.2 Confidence ratings

Confidence ratings were elicited from participants in experiment 3A, and these questions were retained for experiments 3B and 4. The intention was to see if these measures would show that participants (carrying out repetitive causal reasoning tasks for a small amount of money from researchers with whom they had no personal contact) seemed to take longer, and / or be less confident, when making non-normative judgements.

Within each scenario, after having been asked for their single judgement in the change response mode, or the second of their judgements in the delta response mode, participants were asked for their confidence in their judgement. For the change response mode, participants were reminded, on a new screen, of their judgement choice (Less likely / Equally likely / More likely), and asked 'How confident are you in that judgement?'. Responses were by means of a slider control, with the end points marked 'Completely NOT Confident' and 'Completely Confident'. For the delta mode, the only difference was that participants were reminded of the value of the judgement they had just given (for R2).

In addition, for experiments 3B and 4 only, the time participants spent on the screen asking for their judgement (or on each of the two screens for the delta response mode) was recorded. This was considered as a possible proxy for the cognitive difficulty participants experienced in making their judgements. These timings were intended to be collected for experiment 3A also, but a mistake in designing the experiment in Crowdflower meant that no timings were recorded.

These data were used as a fixed factor with delta mode responses, and their interaction, to predict change mode responses in multilevel models with scenario and participant as random factors. Our aim was to assess whether participant confidence in their responses was helpful in predicting the response mode discrepancy.

Two models were fit predicting change ratings from delta ratings for all responses, once with confidence in the delta rating as a predictor, and once without. The LOOIC values for the models with and without the inclusion of confidence as a predictor are shown in table 10. In two cases, the model excluding confidence ratings as a factor was shown to be preferable. For experiment 4, the model including confidence as a factor was slightly preferred.

Table 10

Models of Response Mode Consistency With and Without Participant Rating of Confidence in their Judgements as a Factor

Model	Experiment	WAIC	WAIC weight	LOOIC	Bayesian stacking
Without an effect of confidence	**3A**	**840.90**	**0.93**	**842.51**	**0.92**
With an effect of confidence	3A	846.24	0.07	850.24	0.08
Without an effect of confidence	**3B**	**758.97**	**0.92**	**761.05**	**1**
With an effect of confidence	3B	763.95	0.08	767.79	0
Without an effect of confidence	**4**	**799.85**	**0.08**	**802.34**	**0.44**
With an effect of confidence	4	794.88	0.92	801.10	0.56

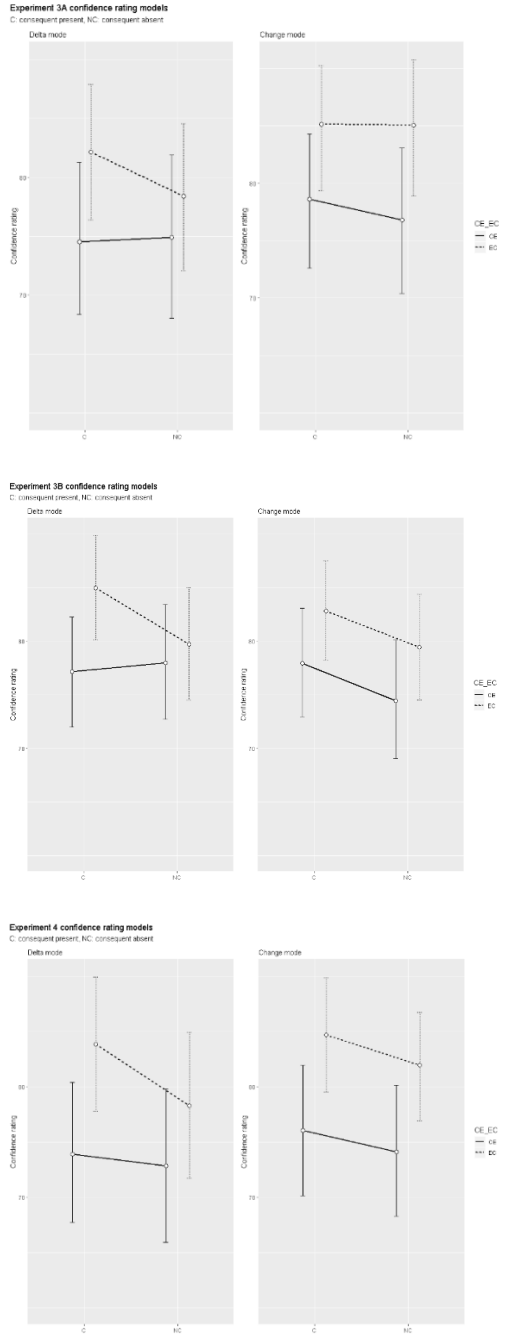
Models of response mode consistency with and without participant rating of confidence in their judgements as a factor. Preferred models shown in bold.

Multi-level models were fit predicting confidence from experimental condition. That is to say, these were models as fit for the discounting / augmenting replications, but including an interaction term, and with confidence rating as the dependent variable, in place of the judgement ratings.

These models are shown in figure 19, for both response modes.

Figure 19

Confidence predicted from condition, with scenario and participant as random effects, for experiments 3A, 3B, and 4



The values predicted by the models for the delta confidence ratings [95% CI] for each condition are, for experiment 3A, CE-C 74.5 [68.4 81.3], CE-NC 74.9 [68.1 81.9], EC-C 82.2 [76.4 87.9], EC-NC 78.4 [72.1 84.5], for experiment 3B, CE-C 77.2 [72.0 82.3], CE-NC 78.0 [72.7 83.4], EC-C 85.0 [80.1 89.9], EC-NC 79.7 [74.5 85.0], and for experiment 4, CE-C 73.9 [67.7 80.4], CE-NC 72.8 [65.9 79.8], EC-C 83.8 [77.8 89.9], EC-NC 78.3 [71.7 84.9].

For the change confidence ratings, the values [95% CI] are, experiment 3A CE-C 78.9 [72.6 84.3], CE-NC 76.8 [70.4 83.1], EC-C 85.1 [79.4 90.2], EC-NC 85.1 [78.9 90.8], experiment 3B CE-C 77.9 [72.9 83.1], CE-NC 74.4 [69.1 80.1], EC-C 82.8 [78.2 87.5], EC-NC 79.7 [74.5 85.0], and for experiment 4, CE-C 76.0 [70.1 81.9], CE-NC 74.1 [68.3 80.1], EC-C 84.7 [79.5 89.8], EC-NC 81.9 [76.9 86.7].

12.2.3 Response times

The response times that participants spent on the two screens where they read the scenario and made their judgement (change mode) or read the scenario and made their judgements (for R1 and R2; delta mode) were treated as reaction times, and trimmed before being used for analyses.

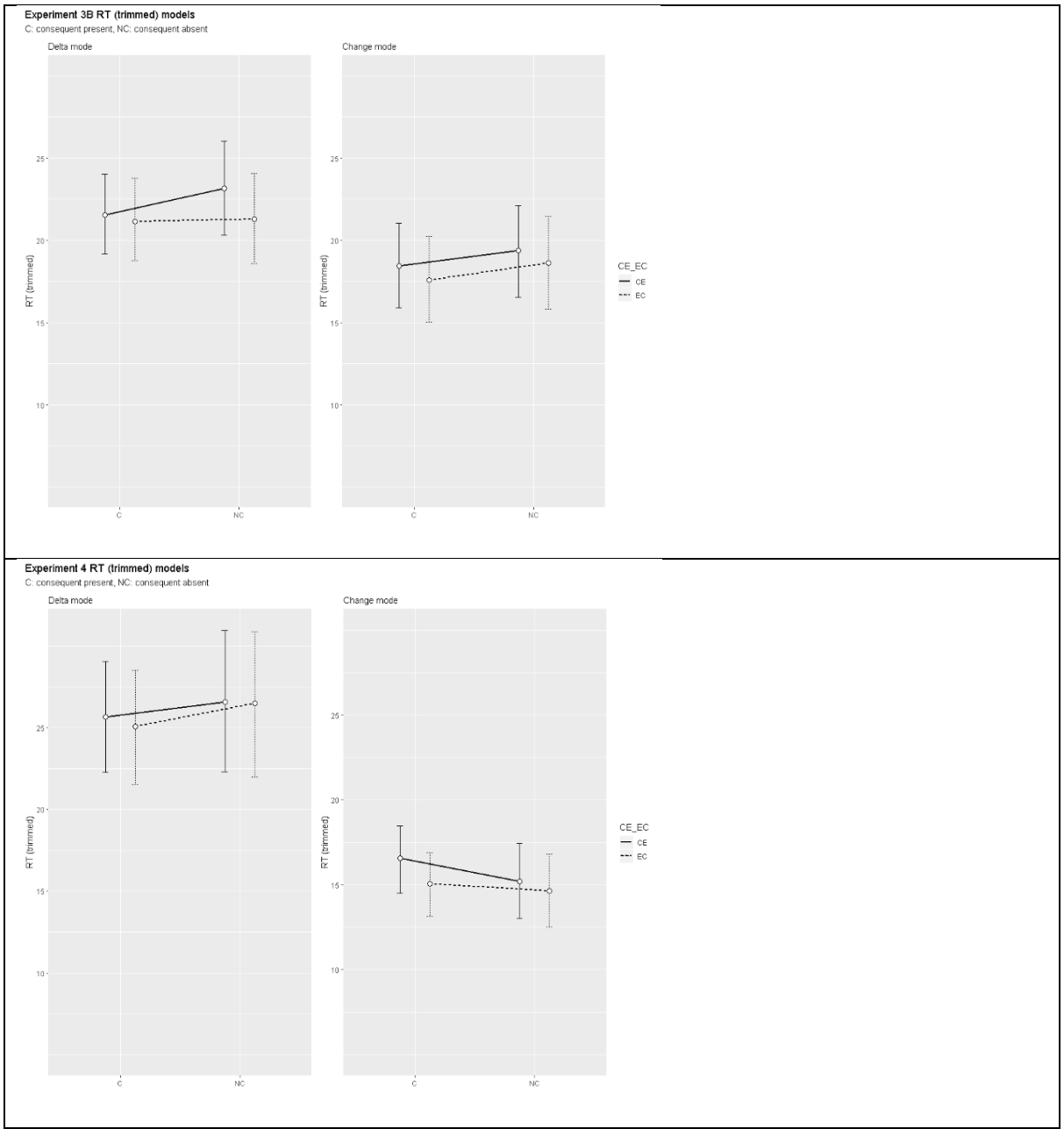
The trimming data were produced by removing all observations above a value of the mean plus twice the standard deviation, and repeating until no more observations were removed. This was done separately for change and for delta. For the change timings, this reduced the experiment 3B observations from 504 to 409, and experiment 4 from 504 to 350. For the delta timings, the experiment 3B observations were reduced from 504 to 344, and experiment 4 from 504 to 415.

Once again, multi-level models were fit predicting response times from experimental condition. That is to say, these were models as fit for the discounting / augmenting replications, but including an interaction term, and with confidence rating as the dependent variable, in place of the judgement ratings.

These models are shown in figure 20, for both response modes.

Figure 20

Response times predicted from condition, with scenario and participant as random effects, for experiments 3B and 4



The values predicted by the models for the delta response timings [95% CI] for each condition are, for experiment 3B, CE-C 21.5 [19.2 24.0], CE-NC 23.1 [20.3 26.0], EC-C 21.1 [18.8 23.8], EC-NC 21.3 [18.6 24.1], and for experiment 4, CE-C 25.7 [22.3 29.1], CE-NC 26.6 [22.3 31.0], EC-C 25.1 [21.5 28.5], EC-NC 26.5 [22.0 30.9].

For the change response timings, the values [95% CI] are, for experiment 3B, CE-C 18.5 [15.9 21.0], CE-NC 19.4 [16.5 22.1], EC-C 17.6 [15.0 20.2], and EC-NC 18.6 [15.8 21.5], and for experiment 4, CE-C 16.6 [14.5 18.5], CE-NC 15.2 [13.0 17.4], EC-C 15.1 [13.1 16.9], and EC-NC 14.7 [12.5 16.8].

12.2.4 Discussion

The credible intervals for the conditions overlap to such an extent that no conclusions can be drawn as to whether the participants were less confident for the conditions in which they reasoned non-normatively. It is unfortunate that the confidence and response time measures were not included in all of the experiments. It seems this question can only be pursued with a study using a considerably larger number of participants. This part of the research cannot be considered a success.

12.3 Phi, a measure of correlation

As described above, a current trend in causal reasoning research is inferentialist, stressing the need for (real or perceived) link between a cause and effect which goes beyond a mere statistical association. We judge the strength of the relationship between 'mud' and 'rain' from how often, or not, we see them together, while the causal direction of the relationship cannot be deduced from frequencies (Pearl, 2001, 2018). Whatever the status of the link, a statistical relationship, a covariation, is necessary to learn and quantify the strength of the link

between cause and effect. In CBN terms, a causal network requires probabilities as well as structure.

There is no single way to quantify covariation. Many different metrics are described and reviewed in Hattori and Oaksford (2007; table 2 gives the formulas for 41 of these measures), among which they take the phi-coefficient as a normative standard to which to compare others. The intention in the research described here was to collect data in phase 3 of the experiments to enable the calculation of phi, and thus to examine if the strength of the relationship between the variables in the causal conditionals was related to the discounting and augmenting in participants' judgements. Various problems intervened and made achieving this goal more difficult than expected.

The phi-coefficient varies from -1, completely negative correlation, for two uncorrelated variables to +1, completely positive correlation, for two variables always seen together. The coefficient can be calculated from a contingency table of the occurrences of two variable - in the present case the two antecedents of the conditional statements. The formula is

$$\phi = \frac{ad-bc}{\sqrt{efgh}}$$

where the terms refer to a contingency table where the cells are labelled as follows:

	P1	¬P1	
P2	a	b	e (= a+b)
¬P2	c	d	f (= c+d)
	g (= a+c)	h (= b+d)	Grand total = N

$$\frac{ad - bc}{\sqrt{(a+b) \times (c+d) \times (a+c) \times (b+d)}}$$

where A,B,C, and D are the 4 cells of the contingency table

In phase 3, for each scenario, the following values were asked of participants:

Question 1: *Pr*(Q|P1)

Question 2: *Pr*(Q|P2)

Question 3: *Pr*(Q|¬P1)

Question 4: *Pr*(Q|¬P2)

Question 5: *Pr*(P2|P1)

Question 6: *Pr*(P1|P2)

Question 7: *Pr*(P1 & P2)

According to probability theory (the definition of conditional probability):

Pr(P1) = *Pr*(P1 & P2)/*Pr*(P2|P1) (i.e. Question 7 / Question 5)

Pr(P2) = *Pr*(P1 & P2)/*Pr*(P1|P2) (i.e. Question 7 / Question 6)

Then,

A = *Pr*(P1 & P2) (i.e. Question 7)

B = Pr(P1 & ¬P2) = *Pr*(P1) - *Pr*(P1 & P2) (i.e. (Question 7 / Question 5) – Question 7)

C = *Pr*(¬P1 & P2) = *Pr*(P2) - *Pr*(P1 & P2) (i.e. (Question 7 / Question 6) – Question 7)

D = *Pr*(¬P1 & ¬P2) = 1 – (*Pr*(P1 & P2) + *Pr*(P1 & ¬P2) + *Pr*(¬P1 & P2)) (i.e. 1 – (A+B+C))

12.3.1 Problems with calculating phi

However, there were difficulties carrying out these calculations. As described above, the responses of participants in the first experiment (and its subsequent replication, i.e. experiments 1A and 1B) were giving by choosing one of seven values to each conditional / conjunctive probability question. E.g.

If she throws the vase, then the vase breaks. If a tennis ball hits the vase, then the vase breaks. How likely is it that she throws the vase AND a tennis ball hits the vase?

'Definitely not', 'Highly unlikely', 'Unlikely', 'Perhaps / Perhaps not', 'Likely', 'Highly likely', 'Definitely yes'

These answers were coded as ranging from 1 to 7, and then put on a probability scale of 0 to 1. If Question 5 or Question 6 was given a value of 'Definitely not', the calculation of *Pr*(P1) or *Pr*(P2), respectively, would fail due to division by zero. In other cases, one of the four values ($(A+B)$, $(C+D)$, $(A+C)$, $(B+D)$) could be zero, leading once again to division by zero for the final calculation.

However, a greater number of problematic cases were due to participants giving answers which were not compatible with the axioms of probability. If the answer for Question 7 was greater than that for either Question 5 or Question 6, this was the case: *Pr*(x & y) cannot be larger than *Pr* (x | y) or *Pr*(y| x). After removing such cases, and those involving division by zero, few judgements remained.

For experiment 2A, and its replication, 2B, the change was made to the use of a visual indication of probability, a horizontal line anchored at 0 and 1, along which participants dragged a slider control to indicate 'how likely'. (At the same time, the way of responding for

the questions in phases 1 and 4 was also changed to this method.) To reduce cases still occurring where participants gave extreme responses leading to division by zero, as described above, the responses were rescaled to be limited within a range of 0.1 to 0.9, replacing cases where participants gave judgements of certainty (in either direction), However, cases of probabilistic incoherence in responding (i.e. the conjunctive probability of the antecedents was higher than that of one of the conditional probabilities) remained high. A further problem also came to light at this point – the questions for the EC / common cause scenarios had been incorrectly written so as to match those for CE. Thus they followed the logical form, and did not take account of the change in causal direction. For the formula of phi, this was not critical, as the calculation was carried out using Question 5, Question 6, and Question 7, which were the same regardless of causal direction. Nonetheless, these questions were rewritten for experiment 3A, and subsequent experiments.

The main change for experiment 3A was a rewriting of the way in which the conjunctive probability was asked, with the aim of making it impossible for participants to give probabilistically incoherent answers.

12.3.2 An improved way of asking the conjunctive probability

An example of the new method is shown for the scenario 'vase' in the delta mode. Participants were reminded of the two conditionals, then asked to give the probability of each antecedent 'regardless' of the other.

If she throws the vase, then the vase breaks.

If a tennis ball hits the vase, then the vase breaks.

Regardless of whether the vase breaks, how likely is it that she throws the vase?

Regardless of whether the vase breaks, how likely is it that a tennis ball hits the vase?

Give your answers by moving the sliders. A value of 0, at the extreme left, means 'definitely not', a value of 0.5, in the middle, means 'perhaps / perhaps not', and a value of 1, at the extreme right, means 'definitely yes'.

Next, participants moved a new screen, which included the less likely (or first, if equal) of the two antecedents just asked, and asked to give the likelihood that both antecedents occur 'as a proportion of that value':

Now we want to ask you about how likely it is that, regardless of whether the vase breaks, she throws the vase AND a tennis ball hits the vase.

We already know that the likelihood that both happen can't be more than the likelihood of one or the other happening.

You have just decided that the lesser likelihood is '[antecedent which was just given lower value]' . We want you to choose the proportion of that value that represents both occurrences happening together.

Move the slider to indicate how likely it is that 'she throws the vase AND a tennis ball hits the vase' compared to '[antecedent just given lower value]'. At the left, 0 means they will 'definitely not' happen together, at the right, 1 means that if '[antecedent just given lower value]', then both will 'definitely yes' happen together, and in the middle one-half (0.5) means that both happening together is half as likely as '[antecedent just given lower value]'.

With this form of questioning, no cases needed to be eliminated due to probabilistic incoherence, and phi could be successfully calculated for experiments 2B and 4. Unfortunately, an error in preparing the conditional questions for experiment 3B meant that it was not possible

to calculate phi for that experiment. Unfortunately, the only experiment for which this analysis could be carried out was experiment 4.

It should be noted here that, in these experiments, where the experimental materials used the terms 'likelihood' it was used as in colloquial language, as a synonym for 'probability'.

12.3.3 Comparison of models fit with and without an effect of phi

Multilevel regression models were fit predicting the delta and change ratings from the causal direction and whether the consequent was asserted, with and without a fixed effect of phi. As before, model comparisons were made using LOOICs (and WAICs for comparison) to show preferred models. Models were fit without an interaction term – this was done to match the preferred models for experiment 4 according to the model comparisons described above (see table 11).

1] A model including phi as a fixed effect

Change Rating / Delta Rating ~ Causal Direction+Assertion of Consequent + Phi + (Causal Direction+Assertion of Consequent + Phi | Participant) + (Assertion of Consequent + Phi | Scenario)

2] A model not including phi as a fixed effect

Change Rating / Delta Rating ~ Causal Direction+Assertion of Consequent + (Causal Direction+Assertion of Consequent | Participant) + (Assertion of Consequent | Scenario)

The delta models originally run in brms for 2,000 iterations failed to give sufficient ESS (effective sample size) values, but succeeded when re-run for 8,000 iterations. The modes are shown in figure 21.

Table 11

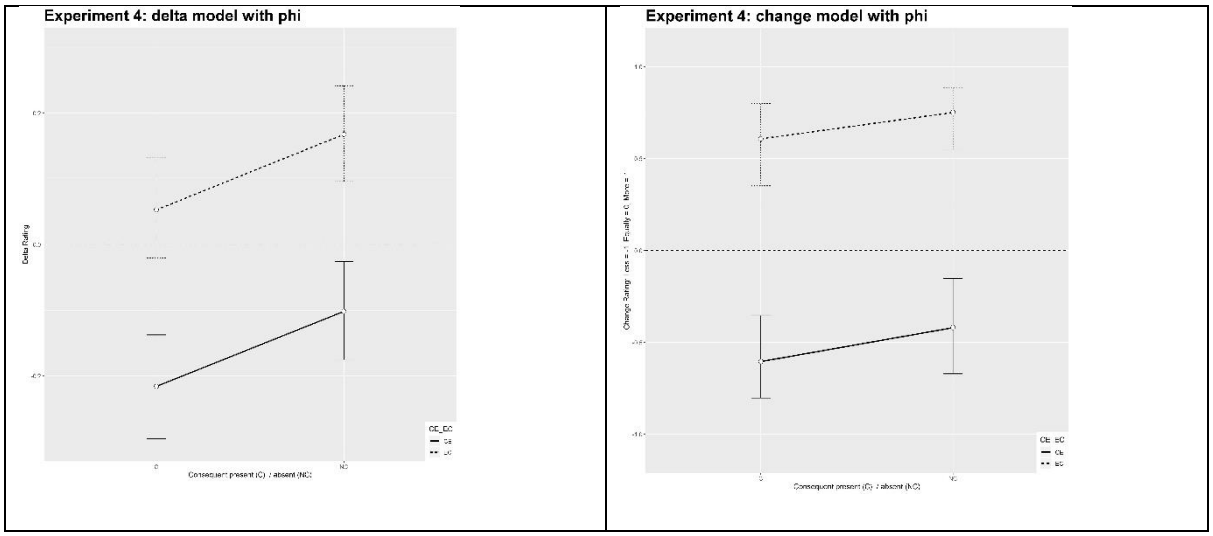
Comparisons for Models With and Without an Effect of Phi

Model	Response mode	WAIC	WAIC weight	LOOIC	Bayesian stacking
Without an effect of phi	**Delta**	**-71.56**	**0.93**	**-69.21**	**1**
With an effect of phi	Delta	-66.30	0.07	-63.24	0
Without an effect of phi	Change	775.24	0	779.48	0.23
With an effect of phi	**Change**	**759.46**	**1**	**766.78**	**0.77**

Comparisons for models with and without an effect of phi. Preferred models in bold.

Figure 21

Models fit for each response mode including phi as a fixed effect.



12.3.4 Discussion

How to interpret this analysis for experiment 4 is unclear. With more data it would have been possible to look for an effect of phi by scenario. For some scenarios, a correlation of the antecedents seems unlikely. For example, for the 'vase' scenario, throwing the vase and also hitting it with a tennis ball would seem to require determination and coordination. That the model comparisons suggest correlation is important for the change mode, but not for the delta mode, suggests that it might have been fruitful if these data had been collected successfully for the other experiments.

12.4 Do participants' conditional probability judgements predict discounting and augmenting?

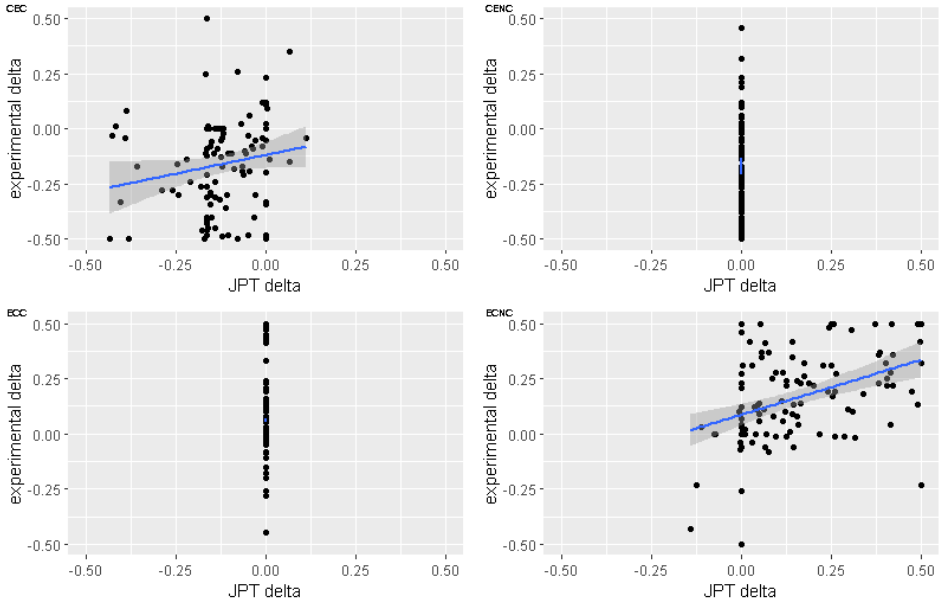
The information in a CBN, or a non-causal Bayes Net, can be expressed as a joint probability table, listing the state of each variable for the state of all the other variables. Beyond simple networks, producing such a table from a CBN is computationally intractable. Calculating a JPT can be made easier by finding all the variables whose states need not be considered with respect to a specific particular variable – these are revealed by the Markov Blanket. Heuristic algorithms can make the remaining calculations simpler. For a three-node network, such as the ones used in the present research, a complete calculation is possible with pen-and-paper, or with a computer. That reasoners faced with such causal relations can use the JPT seems unlikely, but such a representation is normative for comparing their judgements to.

The per-scenario conditional (and conjunctive) probabilities which were collected in phase three of the experiments were also intended to be used for the calculation of the joint probability table (JPT). However, a lack of a value for the probability of the consequent, *Pr*(Q), meant that the calculation could only be completed for experiment 4, for which a question asking for that value was added.

Thus, values can be obtained, on the basis of participants' assessment of each scenario, corresponding to the two ratings (R1 and R2) given during the delta response mode. Subtracting R1 from R2 gives a value that should, normatively, correspond to the delta rating. The delta rating predicted by the JPT values was modelled, with scenario and participant as random factors, using brms. The 95% CI of the population level effect included zero (0.00, [-0.43 0.44]), showing that the JPT is not a good predictor of the delta rating. Figure 22 shows scatterplots of the two values for the two causal directions (CE and EC). For CE-NC and EC-C conditions, the calculated results show neither discounting or augmenting. For CE-C and EC-NC, the association is clear, but weak.

Figure 22

Plots of experiment delta mode responses against delta values calculated from conditional probabilities (Experiment 4)



These results are in line with the model results reported earlier, showing discounting and augmenting where it is normative. In each case the judgements that participants made are more conservative than those to be expected from their understanding of the strength of the

inter-variable associations revealed by their responses to the conditional questions in phase 3. It is again unfortunate that failures in these materials mean that these comparisons can only be made for one of the seven experiments.

Chapter 13 General discussion

13.1 Discounting and augmenting in causal conditional reasoning

The research reported here has confirmed that reasoners about causal conditionals take account of causal dependencies, and do not only take account of the logical form of the conditionals alone. The results for the two causal directions, CE and EC, differ in each response mode. For the change mode, common effect conditional pairs (CE) produced discounting whether the consequent was stated or not. Common cause conditional pairs (EC) produced augmenting regardless of whether the consequent was stated. For the delta mode, the results also differed by causal direction. When the conditional was asserted, the CE condition produced discounting, while the EC condition produced neither discounting or augmenting (normative behaviour for this condition). When the consequent was not asserted, augmenting was produced in the EC condition. For the CE-NC condition, the results were not conclusive, and the normative behaviour of no discounting and no augmenting was not confirmed. However, inspection of the model distribution showed that the tendency was not towards augmenting, and thus once again, the results differed between the two causal directions.

These results are hardly surprising, but they confirm that people take account of causality in conditional reasoning. These results do not, of course, fit the traditional logic-based account of conditional reasoning. They are also not in line with 'traditional' versions of mental model theory. How far they can they are compatible the newer version of MMT (Johnson-Laird et al., 2015) is unclear, as the newer version itself seems not fully worked out, and something of a moving target (Baratgin et al., 2015; Cruz, 2018; Hall et al., 2016; Over & Cruz, 2018).

A widely used and robust conceptual framework based on conditional probabilities for which causal direction is central, and is thus compatible with this finding is the one described above: CBNs.

The problem for a CBN-based account of human reasoning is that people's judgements are often not normative by that account (Derringer & Rottman, 2018a; Rottman & Hastie, 2014). The results of the research reported here also show that reasoners' judgements did not always conform to the predictions of a causal account. Discounting and augmenting were found when predicted – the problem is that discounting and augmenting (or behaviour resembling discounting and augmenting) were found also when not appropriate.

Earlier research has found such non-normative behaviour, often described as violations of the Markov Assumption, when a CBN approach is assumed. For the more-often researched phenomenon of discounting (also called explaining away) earlier research has variously found quantitatively insufficient discounting, no discounting, or augmenting-like behaviour (Rehder & Waldmann, 2017). Rottman and Hastie (2016) found behaviour to be generally qualitatively correct (that is to say, in the normative direction) with some violations. They provided and discussed 12 explanations for non-normative behaviour. Thus, setting aside the response mode for the moment, precedents and possible explanations for the results of the present research, where they are not normative, could be found in earlier research.

13.2 Replicating the response mode discrepancy

However, the central aim of the experiments here was to attempt to replicate the response mode discrepancy, found and addressed in the earlier research of Ali et al., (2010), Ali et al., (2011), and Hall et al., (2016). The discrepancy was replicated, across all seven experiments. At the same time, a discrepancy-within-the-discrepancy, that is to say, the differing status of the EC-C condition in the delta mode, was resolved. Ali et al., (2011) found that participants discounted in this condition, i.e., for common cause conditionals, where discounting is never normative. The present research does not support this result. It is in line

with Ali et al., (2010) where EC-C in the delta mode produced neither discounting nor augmenting.

The response mode discrepancy is important, if it is important, because on the face of it, it is a result that cannot be explained with reference to causality. The causal scenarios in both modes were the same. The importance of the discrepancy is supported by the successful replications reported here.

Attempts to explain the discrepancy began as it was first reported, in Ali et al., (2011). The authors referred first, as is common in discussions of violations of the MA, to the possibility that reasoners were led to consider knowledge of correlations in the real world by the conversational pragmatics of the task: if one of two effects only is stated, reasoners might assume that the experimenters were implicitly suggesting that the other had not occurred. The authors then compared the results in their experiments to those from Ali et al., (2010), where this result was not seen, and suggested that the difference between discounting (2011) and normative behaviour (2010) in the delta mode could be due to the child participants in 2010 performing better as their relative lack of world knowledge and ability to follow conversational pragmatics led them to stick more closely to the scenarios as presented. The present results make such an explanation somewhat less cogent, as the adult participants gave judgements in line with those of the children in the 2010 research.

Hall et al., (2016) examined the 'shallow encoding' hypothesis, suggesting that the higher WM load imposed by the change task might lead to associative, rather than causal reasoning, for the change mode. The hypothesis was not supported, and the authors were faced with the difficulty of giving an account of the discrepancy based on real-world knowledge (correlations) which, of necessity, must be present for both modes: 'if augmentation in the ECC condition using the change ratings is to be explained by an association […] then one would

expect augmentation to also be present using the difference ratings [delta ratings] which it was not' (Hall et al. ,2016, p. 20).

Oaksford and Chater (2017) suggested that the response mode discrepancy in Ali et al., (2011) could be seen as reduced normativity for the change mode (1 condition non normative for delta, 2 non-normative for change) due to the higher WM load imposed by the change mode, requiring storage of a network state while considering a second state before responding. The present results (section 12.2) show a positive correlation between WM score and normativity. This is to be expected for CBNs and for MMT. However, the trend held for both response modes, and was in fact somewhat stronger for the delta mode, which the account in Oaksford and Chater (2017) suggests should require less use of WM.

A different rationale lay behind the considerable changes made to the presentation of the materials for the change mode in the experiment 4 in the current research. Where Hall et al., (2016) considered real-world knowledge as a factor, with the difference in response mode affecting whether that knowledge was used causally (in the delta mode) or merely probabilistically (in the change mode), experiment 4 (section 10.4.2) was designed to examine the possibility that the way the scenarios were presented, not the judgement asked for, or real-world knowledge, was to some extent responsible for the discrepancy.

The change mode presentation was split into two screens, matching that of the delta mode, and the second screen recapitulated the information already given. If the discrepancy had been removed or diminished, that would have been taken as evidence that the response mode discrepancy was of the nature of an experimental artefact: not without interest, but not of importance for understanding causal conditional reasoning. However, despite the changes for experiment 4, the results were in line with those of the previous 6 experiments, and we are left in need of another explanation.

A speculative account follows, introduced by a closer look at discussions in earlier research of the “cue consistency” (Derringer & Rottman, 2018a) first discussed above (section 7.4).

13.3 Speculations on reasons for the discrepancy

If the response mode discrepancy is to be explained, we need to identify how reasoning about the same scenarios can lead to different results for the delta and for the change modes.

To account for the response mode bias, it is obviously necessary to identify a way in which reasoning about the same scenarios may differ between the delta and the change modes. A bias, leading to non-normative reasoning, or bias-like behaviour, explained as normative, or adaptive, by a standard different from that implied by the stated conditionals, will be of no help in explaining the response mode discrepancy, if it operates equally across both modes.

A proposal that the heavier memory load imposed by the change mode leads to a less complex representation of the scenarios, the 'shallow encoding' proposal was not supported by Experiment 2 of Hall et al., (2016). Here I will discuss two areas that may be worth examining for such a difference. Neither proposal is given as a complete, testable, hypothesis: rather, here are preliminary examinations of possible mechanism.

The first proposal is due to Mike Oaksford, considers what might be the effect if the extra model, or its result, needing to be stored in working memory for the change mode, is conducive to a different number of models, or model states, being constructed mentally in comparison for those produced for the delta mode.

The second proposal is based on Tešić et al., (2020), who note the superior explaining away performance of their participants when giving quantitative responses as compared to when they were giving qualitative responses. In the terminology of the present research, the participants produced a more normative pattern of discounting in the delta mode than in the change mode - this is in line with the present research. Tešić et al., (2020), however, note that

the phenomenon of participants, when asked for conditional probabilities, of giving the base rate / unconditional / prior probabilities of the causes which they were supplied with could lead to apparently normative responses for discounting, although such a pattern is grossly non-normative. Thus Tešić et al., (2020) propose that the change mode can be more revealing of participants' performance: "This result highlights the importance of also including qualitative relational questions in such contexts." (Tešić et al., (2020) p. 15). Here the change mode is less normative, but also a better expression of participant's reasoning strategies.

In the discussion of language effects below, I will reject Tešić et al., (2020)'s 'propensity interpretation' hypothesis, but look for a different explanation for participants' responses, and suggest that language effects, reinforcing scenario differences, may be an alternative area to seek for a cause of the response mode discrepancy.

13.3.1 The change mode and mental representation of network states

Is there a way in which the more complex memory task which the change mode requires might lead to more biased (or more 'biased') conditional reasoning?

A proposal of this sort will next be described.

The suggestion will be that it is plausible to consider that the change mode may encourage reasoners to represent more models relating to the causal scenarios they are considering than does the delta mode. Considering more, or fewer, models, for the same reasoning task is clearly reminiscent of the central mechanism of MMT. In the present case, the models will be of two kinds, differing by network structure or by network state. On the one hand, network states which are not merely two required ones, querying one antecedent when the other has or has not been stated to be the case, with the consequent also stated, according

to the C/NC condition. On the other hand, networks going beyond the simple three-node CBNs that are sufficient to represent the pairs of conditionals given to reasoners.

The change mode task differs fundamentally from the delta mode by requiring a comparison of two network states - one value of the variable (P2) in question, a cause in the common effect scenarios (CE), or an effect in the common cause scenarios (EC), must be compared to a second value of the same variable. How this is done is not clear. Reasoners may hold the network, or its joint probability table, in both states in their minds to make the comparison, or they may store a value of the target variable (P2), then dismiss the network before re-representing it in the second state to obtain the necessary second value of P2 to make the comparison, and answer the question: is P2 now more, equally, or less likely? For the delta mode, participants do not need to store anything after making their first response (i.e., supplying R1) before moving on to their second response (supplying their value for R2).

The proposal here is that the necessity to work with two states of a network (or at a minimum to hold a queried value for P2 in short-term memory while moving on to the second state) predisposes reasoners to consider more states, or (more) alternative networks, than they tend to do when carrying out the delta mode task. Supposing this, we can consider each type of additional model, and its effects on discounting / augmenting behaviour, in turn. The examples given here are intended merely to demonstrate the possibilities of such an approach – the values of the variables are not empirically based.

13.3.1 Cue consistency and preferred network states

As described above (7.4), accounts have been given of causal reasoning based on MCMC-style sampling using sufficiently small sample numbers, constraints on the start state for sampling, and constraints on what transitions are allowed, so as to approximate some of the biases observed in experimental tasks. Rehder (2014) used real-world constructs presented in three-node causal networks where the relationships were counterbalanced to reduce the impact

of pre-existing knowledge. The network structures were chain, common-cause, and common-effect; the scenarios described related to the domains of economics, sociology, and meteorology. Participants sometimes reasoned as if variables which were normatively independent were in fact associated. For common-effect scenarios, participants were less likely to carry out discounting than to reason that one cause predicted the other. Using cluster analysis, Rehder divided participants into two groups, with 29% making up a cluster he called "associative reasoners". Such a division suggests the well-known division of human reasoning into Kahneman's two systems, one conscious, slow, and effortful, one fast, effortless and associative (Kahneman, 2011). Experiment 2 in Rehder (2014) gave participants a time deadline to carry out their reasoning tasks, intended to encourage participants to use associative reasoning. This experimental manipulation did not have the intended effect.

Rottman and Hastie (2016) taught participants probabilities for simple three-node causal networks (chain, common-cause, common-effect) and asked them to make inferences on the state of unknown variables. They found that judgements were mostly qualitatively normative, but with some violations. Rottman and Hastie described various explanations for non-normative behaviour, but were not able to choose a best explanation, or combination of explanations, for their results. What Rottman and Hastie call the "monotonicity assumption", which they proposed might lie behind a number of their results, is similar to Rehder's (2014) associative bias: "judgements tended to be monotonically related to the number of cues that are present minus the number of cues that are absent" (Rottman & Hastie, 2016, p. 120).

In Rehder and Waldmann (2017) the "associative bias" put forward in Rehder (2014) is called "the rich-get-richer principle". Rehder and Waldmann carried out experiments to examine reasoning about common-effect and common-cause networks presented to participants either as verbal descriptions, or via co-variation data given as simultaneous examples (rather than trial-by-trial). Their prediction was that reasoning performance would

be less normative when statistical data were accompanied by verbal descriptions of a causal scenario than for presentation of statistical data only, and least normative for scenarios merely described, rather than experienced. According to the rich-get-richer principle, description of causal relationships activated causal thinking, and its associated bias, which would lead to poorer reasoning than that based on figures showing statistical associations only, without a causal explanation. (This slightly confusing effect, by which an associative bias is not activated when only data on statistical associations is available, is presumably the reason for the change in the name of the bias from 2014 to 2017.) The results of the experiments in Rehder and Waldmann (2017) were in line with their predictions.

Derringer and Rottman (2018a) used the term "cue consistency" for the type of reasoning described in this section, and looked at whether reasoners have more difficulty with structures leading to augmenting (common-cause and chain, classed together by the authors as mediation structures) or with three-node structures leading to discounting (called explaining-away by the authors, predicted for common-effect scenarios). (Derringer and Rottman's terminology differs also from that used here in that they described only the non-normative behaviour for mediation structures as Markov Assumption violations.) Building on the results of Rehder and Waldmann (2017) that verbal descriptions of scenarios led to worse (i.e., more consistent or associative results) they used statistical learning trials without verbal descriptions.

Derringer and Rottman (2018b) prepared learning trial frequencies such that the predicted probabilities for examples of the two structure types closely matched. This required presenting the common effect structure in two variants (with one of the causes either present or absent, and the other present, along with a present effect), one of which had the probability 'flipped' to match the other case and structure. In addition, to prevent simple 'perceptual matching' of variable states, they counterbalanced the variable valences as had been done previously in Rehder (2014) and in Rehder and Waldmann (2017) (e.g., high / low interest rates

caused high / low savings etc.), such that cue consistency effects would require understanding of the associative / causal relations, in this case between the colour and features of fictional microbes.

Rehder and Waldmann (2018b) replicated the results most prevalent in earlier studies, that is to say, they found less discounting and augmenting than was normative. Comparing across structures, they found that judgements were less normative for one of the common effect variants than for the mediation structure, and more normative for the other common effect variant. Thus, Rehder and Waldmann concluded that reasoners did not inherently reason worse with either of the three-node structures. At the same time, in line with Rehder and Waldmann (2017), they concluded that cue consistency, based on a learning rather than a judgement bias, could explain Markov Assumption violations.

Derringer and Rottman (2018b) carried out a series of experiments which demonstrated that participants learned causal relations with multiple causes of one effect more efficiently when the changes between learning trials took place in one cue (variable) at time, rather than in multiple variables. A similar preference for changes in one variable at a time informs a model, not of causal learning, but of causal inference proposed by Davis and Rehder (2017), based on a modified Markov Chain Monte Carlo sampling method which begins its journey through the sample space with all variables at a consistent value, and proposes a next position which differs in a single variable. I will next introduce MCMC sampling methods.

The ways in which CBNs have been applied to computational modelling of causal relationships is of interest if we wish to consider CBN formalities as throwing light on the algorithmic as well as computational level of human causal reasoning (Marr, 1982). Despite the computational simplification given by sparse CBNs, where absent links signify unconditional independence, and reduce the need for excessively time-intensive calculation of a full joint probability table, the calculations can still be intractable as larger networks are

modelled. Thus, approximate algorithmic methods have been applied (Korb & Nicholson, 2011; Neapolitan, 2004). MCMC modelling can simplify computations dramatically, and allow approximations to be made more accurate by investing more time and computational power in improving the solution. If human reasoners can combine, as needed, knowledge of the causal strengths of pairwise relationships between events in the real world, into small models of causal relationships, mental approximations of the probability distributions implied by the networks might be obtained by throwing the mental representation into different states, and observing (presumably unconsciously) the resulting network state (Barsalou, 1999; Hagmayer & Waldmann, 2000).

Davis and Rehder (2017) proposed a theory of causal reasoning based on limited sampling from mental, causal, network models. Their model, the "mutation sampler", is a variation of the well-known and commonly used MCMC sampling methods, in particular the Metropolis-Hastings version.

Although CBNs can implement both continuous and discrete variables, they are more often used with discrete variables, typically standing for whether an event occurs or not, or whether a condition is present or not. In psychological research into reasoning with causal network, such discrete variables are the norm. The mutation sampler also works with discrete variables. Thus, the three-node networks needed for the simplest cases of common-effect and common-cause scenarios have a total of eight distinct network states.

MCMC samplers move through a sample space along a path determined only by the current position and its predecessor. A characteristic of the mutation sampler is that a proposed new position the path (the chain) can only differ from its predecessor in the value of one variable.

The mutation sampler, as proposed by Davis and Rehder (2020), restricts its starting position in the chain of samples to the two consistent network states: the variables are either

all 0 (e.g., false) or 1 (e.g., true). As with traditional MCMC samplers, the starting position is irrelevant to the accuracy of the results given sufficient sampling time (a long enough chain). Limited sampling is, however, a feature of the mutation sampler, and it is this which allows the sampler to produce some of the same biases found experimentally in human causal reasoning.

13.3.2 States, models, and the response mode discrepancy

These proposals partly inform a possible account of the response mode bias (due to Mike Oaksford), which will now be summarised. This account follows Davis and Rehder (2017) as seeing the proposed constraint on the (consistent) starting state for sampling as similar to the preference in MMT for a state in which the antecedent and consequent of a conditional are most effortlessly represented as consistent (in MMT, both true: Johnson-Laird & Byrne, 2002). Although the various CBN states that participants in the current experiments were required to assess in making their judgements are not the states of a sampling process (they are not a random walk), just as MMT inspired the simplified sampling procedure proposed in Davis and Rehder (2020), so that procedure inspires the proposal here as to what reasoners may be doing in the experiments described here. In particular, the change mode task is seen as leading to constraints in what states (or 'models') are considered. The proposal is that there is a bias towards considering the effects of different cause states, that is to say, reasoners will prefer to consider the different possible states (true or false) of the causes in a network considered, rather than the different states of any effects present, or an alternative value for a variable which has already been judged (or 'queried'; for example, the cause in the second conditional stated in common effect scenarios (CE), or the effect in the second conditional stated in common cause (EC) scenarios). The preference for changing cause states is seen as due to reasoners' preference for simulating in the same direction as time's arrow. i.e., predictively rather than diagnostically.

The present suggestion differs in some ways from the account of sampling in Davis and Rehder (2020), and from earlier suggestions of a 'monotonicity assumption' (Rehder, 2014; Rehder & Waldmann, 2017; Rottman & Hastie, 2016). In the present experiments, participants were asked to consider two causal conditionals, and to judge one of the antecedents on that basis, before being given further information. This 'baseline' judgement took place with the two other variables in an indeterminate state – whether the participants tended to assume all the nodes were true, or all the nodes were false is unclear. The sampling proposal of Davis and Rehder (2020) is of a bias caused by a fixed commencing state, followed by sampling with the likelihood of termination before sufficient samples had been taken to allow for convergence (although a normative result will occur if sampling continues indefinitely). The constraint that the state of only one variable can vary for the next, proposed, state for comparison is a constraint making it less likely that the mirror-image state, in which all variables are consistent, as for the commencement state, but flipped in valency, will be reached when limited samples are taken.

By contrast, the proposal here is for the constraint that the final state considered must differ from its predecessor by the state of a cause. In the account of Davis and Rehder (2020), there is no constraint on the last state considered – in fact, the last state is merely due to when sampling ends. A bias affecting the first configuration of the network considered by reasoners, as in Davis and Rehder, would seem not to help in explaining the response mode discrepancy, inasmuch as for both modes, the 'baseline' condition, due to the presentation of the two conditionals, is the same. In general, the proposal here should be seen as analogous to that of Davis and Rehder (2020), rather than an extension of it, since Davis and Rehder proposed that reasoners sample to produce a joint probability table for the network. Once reasoners have a JPT, further sampling should presumably not be needed until a judgement is given, and the JPT dismissed (at end of the change task, or after giving R1 and R2 for the delta mode).

Nonetheless, assuming that participants tended to complete their judgements by reversing the state of a cause, and that where the task does not demand they consider such a model (network state), they do so anyway, offers one possible explanation for the consistent augmenting for the EC scenarios in the change mode. However, as we will see, in its current form, this proposal describes elevated augmenting behaviour in the change mode EC-NC condition, where augmenting is normative, and present for the delta mode, rather than for the more problematic EC-C condition, where augmenting-like behaviour in the change mode distinguishes it from the delta mode.

The following tables give an overview of the results that would be expected by this account. For these results, for CE, shown in table 12, the two causes have a background probability of 0.5, and (using a noisy-OR function) when one or both causes is present, the effect has a probability of 0.99. In the case where neither cause is true, the effect has a probability of 0.05 (i.e. $Pr(C \mid \neg P1 \ \& \ \neg P2) = 0.05$)).

For EC, shown in table 11, the single cause has a background probability of 0.5, and each effect has a probability of 0.8 when the cause is present, and 0.2 when it is absent.

Table 12

CE Judgements with a Final Change of Cause State

CE-C	Stated / Assumed true	C2? – the queried variable	C1	E	Rating	Change – the stored network state
Model 0	Conditionals only (Baseline)	.50	.50	.76	None (R0)	C1 C2? **E (.50)**
Model 1	**E**	.66 (R1)	.66	E = 1	R1	**C1 C2**? **E (.66)**
Model 2	**E, C1 (P1)**	.50 (R2)	C1 = 1	E = 1	R2	**C1** C2? **E (.50): rating is 'less', since M2 < M1: discounting**
CE-NC	Stated / Assumed true	C2? – the queried variable	C1	E	Rating	Change – the stored network state
M0	Conditionals only (Baseline)	.50 (R1)	.50	.76	R1	C1 C2? **E (.50)**
M1	**C1 (P1)**	.50 (R2)	C1 = 1	.99	R2	**C1** C2? **E (.50)**: rating is 'equally', since M2 = M1: no discounting

For the CE conditions, the final judgement (delta, R2), or the second of the compared values (change) is produced after changing the state of one of the causes.

Table 13

EC Judgements with a Final Change of Cause State

EC-C	Stated / Assumed true	E2? – the queried variable	E1	C	Rating	Change – the stored network state
M0	Baseline	.50	.50	.50	R0	E1 E2? C **(.50)**
M1	**C**	.80	.80	C = 1	R1	**E1 E2**? **C (.80)**
M2	**C, E1 (P1)**	.80	E1 = 1	C = 1	R2	**E1** C2? C **(.80):** rating is 'equally, since M2 = M1: neither
M3	**E1 (P1)**	.80	E1 = 1	.80	R3	**E1** C2? C **(.80):** rating is 'equally', since M2 = M1: neither
EC-NC	Stated / Assumed true	E2? – the queried variable	E1	C	Rating	Change – the stored network state
M0	Baseline	.50	.50	.50	R1	E1 E2? C **(.50)**
M1	**E1 (P1)**	.68	E1 = 1	.80	R2	**E1 E2**? **C (.68):** rating is 'more', since M1 > M0: augmenting
M2	**C, E1 (P1)**	.80	E1 = 1	C = 1	R3	**E1** C2? C **(.80):** rating is 'more', since M2 > M0: enhanced augmenting

For the EC conditions, after the network state needed for the final judgement (delta, R2), or the second of the compared values (change) is produced, an additional state is considered, by reversing the state of the single cause. This extra model, conducive to bias, is shown in grey.

The critical point here is the proposal that participants were more likely to produce the final model shown in table 13 (the line in grey), because the change task is inherently requiring them to hold more than one model (network state) in mind to produce their judgement. This, if it is the case, will be the difference between the response modes that is necessary for an explanation of the discrepancy. The final model will presumably have a greater effect due to its recency, as the last network state represented. In this way, it will be analogous to the starting, 'congruent', network state for sampling as proposed by Davis and Rehder (2020), that is to say, a model which has a larger influence on reasoners' assessments than its peers. This type of effect of an extra model, with a mutation of a cause is, notably, independent of the content of the reasoning scenario. By contrast, the second type of additional model it is proposed that will be selectively entertained more by reasoners in the change mode is entirely content dependent.

It is often proposed that conditional reasoners consider networks different from those presented to participants in reasoning experiments, giving a different standard for what is normative to that proposed by an experimenter, and so revealing a 'bias' as illusory, is one that has very often been put forward (e.g., Hagmayer, 2016; Mayrhofer & Waldmann, 2014; Park & Sloman, 2013). Such proposals are not directly relevant to the response mode discrepancy, if one assumes that reasoners in each mode use their real-world knowledge to create the same network of causal relationships different from that desired by the experimenter. The possible insight that may make this consideration relevant to the response mode discrepancy, is to link the representation of extended, additional, causal models to the need in the change mode to work with more than one network state at a time.

An intuitively convincing example of such an extended model is as follows. Consider the two causal conditionals:

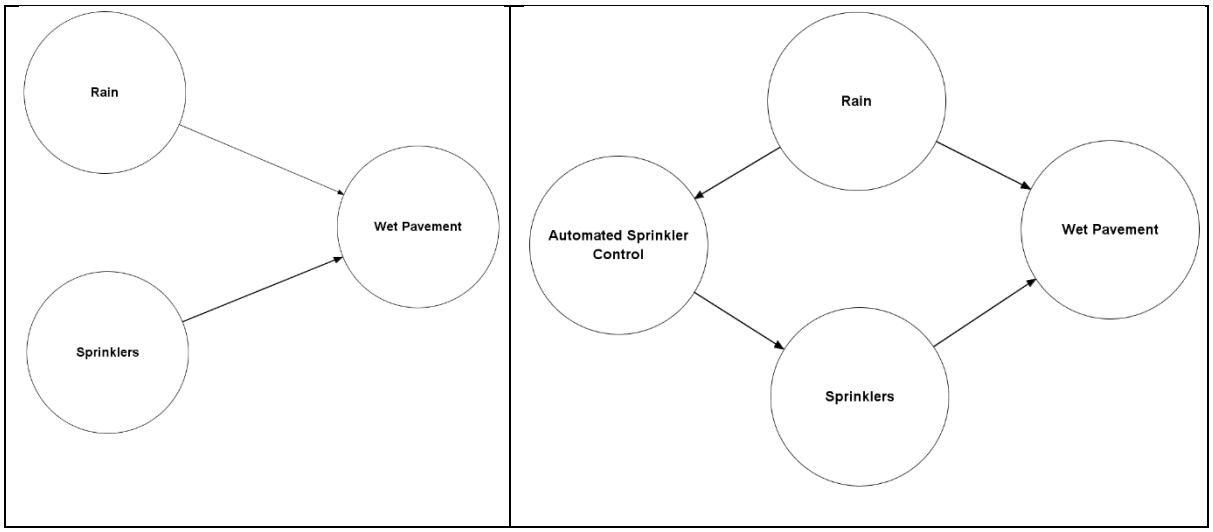
If it is raining, the pavement is wet.

If the sprinklers are turned on, the pavement is wet.

These sentences describe a common-effect scenario. Our real-world knowledge of rain and sprinklers is likely to invite us to consider that gardeners are likely to turn their sprinklers on when the weather is dry, and turn them off if it is raining. If contact with causal reasoning research has led us to be nervous about human agents and their actions as causes and effects in CBNs, we may prefer to consider an automated sprinkler system, the sensors of which are the first link in a mechanism which leads to the sprinkler system shutting off when rain is detected. These two CBNs, that implied by the conditional pair, and that extended on account of our world knowledge are shown in figure 23.

Figure 23

A pair of common-effect CBNS. The network on the right is derived from that on the left with the addition of knowledge of the world.



Once again, we can set plausible values for the associated probabilities to examine the effect of such an extended network. We assume that a lack of rain makes the sprinklers more

likely to be in use, by a causal chain going through the added node (or its human alternative, an alert gardener).

Table 14 shows the results of editing the generic CE model in table 10 by including an additional node A, with a probability of 0.99 if there is rain, and setting the probability of the sprinkler node to be 0.99 if A is active (i.e., Pr(A | rain) = .01, Pr(A | no rain) = .99, Pr(Sprinklers | A) = .99, Pr(Sprinklers | ¬A) = .01).

Table 14

Values from an Extended Sprinkler Network

CE-C	Stated / Assumed true	C2? – the queried variable	C1	E	Rating	Change – the stored network state
M0	Baseline	.50	.50	.98	R0	C1 C2? **E (.50)**
M1	**E**	.50	.50	E = 1	R1	**C1 C2**? **E (.50)**
M2	**E, C1 (P1)**	.02	C1 = 1	E = 1	R2	**C1 ¬C2**? **E (.02) : rating is 'less', since M2 < M1: discounting**

CE-NC	Stated / Assumed true	C2? – the queried variable	C1	E	Rating	Change – the stored network state
M0	Baseline	.50	.50	.98	R1	C1 C2? **E (.50)**
M1	**C1 (P1)**	.02	C1 = 1	.99	R2	**C1 ¬C2**? **E (.02): rating is 'less', since M2 < M1: discounting**

In fact, this particular scenario, ‘wet pavement’, given here because of its intuitive appeal, was not included in the scenarios used in the present research. Common-effect examples from these experiments where one cause being true might seem to make the other more likely to be false are among those used, however.

Scenario ‘stereo’:

If the fuse on the stereo is blown, then the stereo is off.

If the stereo is unplugged, then the stereo is off.

If an electrical appliance is unplugged, it is unlikely that an internal fuse will blow.

Scenario ‘jam’:

If there is an accident on the main road, then he is caught in a traffic jam.

If there are road works on the main road, then he is caught in a traffic jam.

If road works slow down traffic, an accident may seem less likely.

Scenario ‘upset’:

If she is being fired from her job, then she gets upset.

If she is breaking up with her partner, then she gets upset.

If she has been fired, she is more likely to be patient with her partner’s annoying personality, for the sake of support at a difficult time, and her partner may avoid being so inconsiderate as to break up when she has just been fired.

Similar considerations apply to EC (common-effect) scenarios. For example,

If your colleague has time off work, then your colleague is ill.

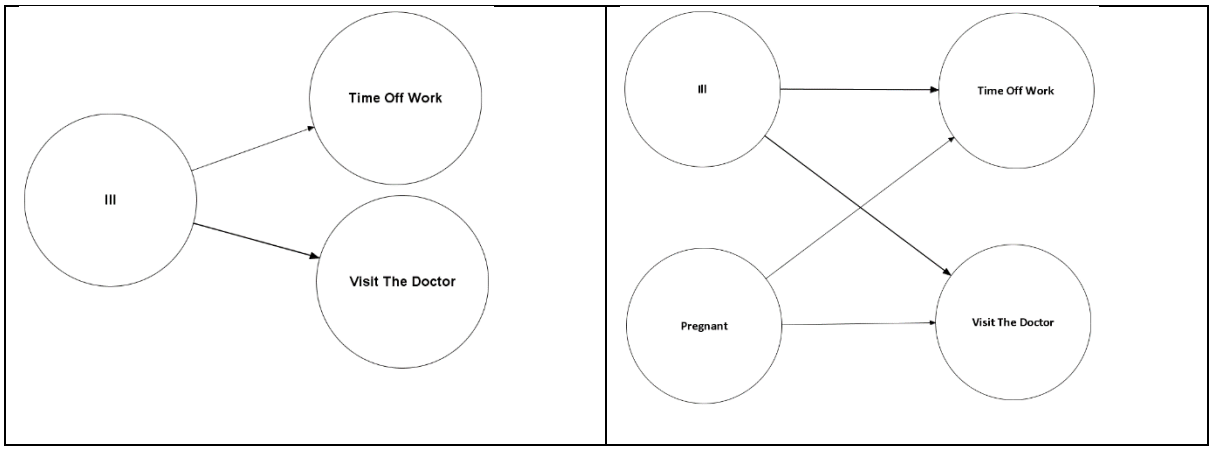
If your colleague visits the doctor, then your colleague is ill.

In such a scenario, we can consider an alternative cause of the two effects. For example, a colleague who is heavily pregnant may also visit the doctor and have time off work.

The two CBNs associated with this conditional pair, as given, and extended by the extra node ‘pregnant’, are shown in figure 24.

Figure 24

A pair of common-cause CBNS. The network on the right is derived from that on the left with the addition of knowledge of the world.



Setting the background probability of both ill and pregnant at .5, and setting *Pr*(either effect | Ill & Pregnant) = *Pr*(either effect | Pregnant) = .99, and *Pr*(either effect | Ill) = .6, and *Pr*(either effect | ¬Ill & ¬ Pregnant) = .01, the values in table 15 are obtained.

Table 15

Values from the Extended Ill Network

EC-C	Stated / Assumed true	E2? – the queried variable	E1	C	Rating	Change – the stored network state
M0	Baseline	.65	.65	.50	R0	E1 E2? C **(.65)**
M1	**C**	.80	.80	C = 1	R1	**E1 E2**? **C (.80)**
M2	**C, E1 (P1)**	.84	E1 = 1	C = 1	R2	**E1** C2? C **(.84):** rating is 'more', since M2 > M1: augmenting
M3	**E1 (P1)**	.98	E1 = 1	C = 0	R3	**E1** C2? C **(.98):** rating is 'more', since M3 > M1: enhanced augmenting

EC-NC	Stated / Assumed true	E2? – the queried variable	E1	C	Rating	Change – the stored network state
M0	Baseline	.65	.65	.50	R1	E1 E2? C **(.65)**
M1	**E1 (P1)**	.90	E1 = 1	.61	R2	**E1 E2**? **C (.90):** rating is 'more',

						since M1 > M0: augmenting
M2	**C, E1 (P1)**	.84	E1 = 1	C = 1	R3	**E1** C2? C **(.91):** rating is 'more', since M2 > M0: diminished augmenting

The results show augmenting for EC-C, made stronger if the final 'flip' of the cause (ill) is made. For EC-NC, this final change of the cause state slightly diminishes augmenting, but the value is still high.

A similar possibility for extending the implied CBN by a node, might apply, for example, to the EC scenarios 'cold gran'.

Cold gran:

If she is shivering, then she feels cold.

If the hairs on her arms are raised, then she feels cold.

Grandma might shiver and the hairs on her arms might be raised because of some medical condition, or some medication (or a recreational drug, if she is of the Woodstock generation). In general, it seems very likely that pairs of what might be described as symptoms could have an alternative cause, in line with the difficulty of diagnoses generally.

Whether participants are considering alternative modes in the sense of network states or network structures, the central point of this proposed explanation is the difference between the modes, due to the requirement of the change mode for considering and holding in memory alternative models. If this indeed led participants to be more likely to represent further models, it may lie behind the response mode discrepancy. So far, of course, this is merely speculative.

A particular connection between the representation by reasoners of different models, and the response mode discrepancy would be given if reasoners are led to consider more, and **extended**, models in the change mode. The rationale behind this proposal is that the change task already imposes the holding of more network states in the mind than does the delta mode. In the delta mode, the first queried model can be disposed of as soon as the response has been given (i.e., as soon as value R1 has been typed in for experiments 1A and 1B, or the slider has been dragged to the desired position, for the remaining experiments).

13.3.3 Questions raised by this proposal

An explanation such as that proposed here is obviously speculative, but the striking consistency of the response mode discrepancy across so many experiments, following an earlier lack of success in elucidating what might lie behind the discrepancy, means that speculation is required.

It is unfortunate that shortcomings in materials and data collection mean that not as much was learned as was hoped in the present research to go beyond the successful replication. Further research might try to rectify these failures, and, since the status of the replication is clear, alter the tasks to try to produce more informative results.

For example, what is the status of the extra memory load imposed by the change mode? What effect would imposing an unconnected memory load have on conditional reasoning of the type used in these experiments? For example, if reasoners in the delta mode were asked to keep unrelated information in mind while completing their task, would the results more closely resemble those for the change mode reported here?

Would careful construction of scenarios so as to make extended networks more or less plausible affect the discrepancy? That is to say, if the change mode memory load encourages representing alternative models, would the discrepancy be reduced where alternative models were less compatible with the real world? The multi-level analyses used for the discounting

and augmenting models in this research were intended to protect the results from unexpected influences due to the particular wording of the materials as also in Hall et al., 2016 (Clark, 1973). However, if the salience of alternative models, which must vary for different non-abstract scenarios, is of central importance, pre-testing materials might be the best approach.

Further, is it certain that reasoners in the delta mode do indeed dismiss the model for the first judgement (R1) from their minds before moving on to a second judgement (R2) on the same topic? If that is indeed the case, would a task requiring the consideration of a different scenario in between those two judgements have no effect on the delta mode results? Or is it the case that, though participants could in theory dispose of a model before creating a new one, for the delta mode, but not for the change mode, in practice, they learn early on that the scenario is not over and done with until R2 has been delivered?

If there is a bias to end consideration of a scenario with an alternative state of a cause, and this has a stronger effect for common cause (EC) scenarios, would judgements be different with materials that did not only have the cause unasserted, or said to be true, but presented the cause also in a 'false' condition. It may be worth considering whether there is a different preference for predictive reasoning, following the arrow of time, and diagnostic reasoning, in the other direction, between physical, billiard-ball style scenarios, and understanding the motivations of conscious agents. This is reflected in the expression 'last in execution, first in intention': "the principle with respect to intention is the ultimate end; the principle with respect to execution is the first means related to the end" (Aquinas, 1983, p. 9). The cause of an automatic-sprinkler system shutting off may be precipitation; but the cause of a gardener shutting of a sprinkler system is likely to be understood as his or her intention with respect to an end state.

13.3.2 Natural language effects, and scenario effects

13.3.2.1 Tešić et al., (2020)'s hypothesis

Tešić et al., (2020) found many of their participants, when asked for conditional probabilities, gave the prior / unconditional probabilities of the causes which they had been supplied with (*Pr*(Ci|E) or *Pr*(Ci|E,Cj) = *Pr*(Ci)). Tešić et al., (2020) asked their participants to explain their reasoning. These responses are not given in the study, but a summary of these participant explanations (relating, however, to the responses for the earlier study, Liefgreen et al., (2018), is given thus:

"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" (Tešić et al., 2020, p. 6).

Tešić et al., (2020) identify two groups among their participants, each with a particular non-normative reasoning strategy: 'diagnostic split strategy', and 'propensity interpretation'. The diagnostic split strategy is viewed by Tešić et al., as a non-normative heuristic, but one which gives judgements which are closer to normality as the base rates of the causes are lower. Tešić et al., (2020) note that Rottman and Hastie (2016) and Pilditch, Fenton, and Lagnado (2019) previously found evidence of participants using similar strategies.

For the group of participants using what they term the 'propensity interpretation', Tešić et al., (2020) do not see their answers as resulting from a bias or heuristic which may at times give normative, or nearly normative results, but rather as evidence that the participants are reasoning in line with a particular understanding of the meaning of probability. Where Liefgreen et al., (2018) used only one scenario, though specified in more than one network configuration, Tešić et al., (2020) introduced new scenarios which they expected would lead participants to use one of the more common interpretations of probability, as a limiting

frequency, or a belief. These three scenarios which Tešić et al., (2020) use are worth examining, but first I will look briefly at the ways in which using a propensity understanding of probability may lead to different reasoning results. The better-known interpretations of probability, as a limiting frequency and a as a metric of belief, do not lead to different normative standards – in both cases, probability values are handled in line with the Kolmogorov axioms.

The suggestion that the third interpretation, propensity, might not be compatible with the Kolmogorov axioms, was first put forward some decades after the propensity interpretation itself. Popper, (1957), cited in (Gillies, 2016), first suggested this interpretation, seeking for a way to give an *objective* probability to *single* events: the two other interpretations cannot do this, as single events do not have limiting frequencies, and beliefs are not objective. Since Popper's proposal the propensity interpretation has become a 'diffuse set of proposals' (Gillies, 2012; Miller, 1994, p. 175). The claim that propensities are not compatible with the Kolmogorov axioms (otherwise stated as 'propensities are not probabilities') is due to Humphreys (1985), who suggested that propensities, seen as probabilistic causes, cannot be applied overall to a system of probabilities compatible with the Kolmogorov axioms, since an effect is not a propensity, and thus diagnostic reasoning cannot be founded upon propensities (Gillies, 2016), This is a problem if one wishes to interpret *all* probabilities as propensities, which does not seem to have been Popper's intention. Alternatives to Kolmogorov, compatible with a particular understanding of propensity, have been suggested, for example, in Fetzer (2012). Aside from the usually accepted justifications for seeing standard probability as normative for human reasoning, for example the 'Dutchbook' arguments, which mean that actually applying such a non-standard interpretation to human reasoning would require starting from scratch in probabilistic reasoning research, it is also important to note that there seems to be no specific alternative non-Kolmogorov framework which requires diagnostic reasoning to give the base rate causal probabilities regardless of the status of the effect, which is what is

needed if Tešić et al., (2020)'s participants are actually using a 'propensity interpretation'. If, on the other hand, in line with the original understanding of propensities, supported by some / most researchers (Gillies, 2016), probabilities interpreted as propensities (not necessarily for every probabilistic relationship) do comply with the traditional framework relying on the Kolmogorov axioms, then the participants' interpretation is not predictive of a different pattern of responses - the answers should be the same for subjectivists, frequentists, and propensity interpreters. (Tešić et al., (2020) also seem to accept that their participants should follow the Kolmogorov axioms when they call the participants' 'propensity interpretation' non-normative.) Thus, the interesting results of Tešić et al., (2020) seem to require explanation as participant error, rather than use of a particular variety of an interesting, but not particularly productive understanding of probability. Nonetheless, the manipulation of scenarios used to test the propensity interpretation was successful. Is the explanation put forward by Tešić et al., (2020) the best? If we consider the two interpretations of probability which dominate, and are widely used, frequentist and subjective, we may think that Tešić et al., (2020) have given their participants scenarios which intuitively lead to one or the other of those two, more well-known, interpretations, and that this has led to a difference in the pattern of errors, with 'experimenter demand', a desire by participants to respond in line with the experimenters (perceived) understanding of the meaning of probability, for the first two scenarios.

If the participants assumed that, since the questions they were asked in the experiments centred around 'probability', the experimenters were interested in how well the participants reasoned about probability, they were certainly correct. Popular science books discuss whether the general public has misconceptions about probability which lead it to reason poorly (Kahneman, Slovic, & Tversky, 1982) or which do not stop it from reasoning well (Gigerenzer & Todd, 1999). Mass media regularly features discussions of everyday failing to understand the language of probability, from weather forecasts, to medical tests, to Covid. The youngish

(mean age = 34.6 year) undergraduate participants in Tešić et al., (2020) are likely to have been exposed in school to educational material stressing the importance of avoiding common misconceptions about probability (Bryant and Nunes, 2012). If participants had the impression that probability is a slippery but important concept that has incorrect common-sense interpretations was the despair of experts, they might give back the priors they had just been told in answer to questions about conditional probability, particularly when the scenarios described bizarre mechanisms apparently created specifically for experimental purposes. The participants might abandon this strategy when faced with a scenario (red wine) which does not resemble those typically found in instructional materials, and which suits less a frequentist interpretation, of a scenario which would conveniently be investigated by a large number of trials.

13.3.2.1 Tešić et al., (2020)'s materials and natural language

The first scenario is of two coins tossed by a mechanism, without human intervention, which lit up a light bulb whenever the two coins were not both tails. This scenario appears in both Liefgreen et al., (2018) and Tešić et al., (2020), and the results which it led to in 2018 are the reason that Tešić et al., (2020) developed their theory of the 'propensity interpretation'. In the first experiment of Tešić et al., (2020), two more scenarios were introduced, one broadly similar, of a mechanism where balls chosen randomly, completed an electrical circuit to again light up a light bulb. The third scenario was specifically developed to test the 'propensity interpretation' hypothesis. This scenario instructed participants to consider a party of which they were the host. No 'mechanism' was involved - the effect was a guest drinking red wine, and the two causes were guests who had been requested to bring red wine. As Tešić et al., (2020) predicted, for this scenario, the proportion of participants who did not update, but rather gave the unconditional probabilities of the causes, ignoring the status of the effect, was smaller. However, these results do not support only the 'propensity interpretation'. Coins, or balls,

flipped, or chosen randomly, are interchangeable entities (thus they lack proper names) - these are trials for which it seems very reasonable to supply matching base rates for different objects - two coins, or two containers of balls. Percentages of 20%, 50$, or 80% are intuitively understandable as obtained by simply carrying out many trials. Who would be confident in assigning a probability – a 'propensity' - of 0.2 for heads to a not-yet flipped coin? These two scenarios are precisely of the sort which cause no difficulties for the frequentist view of probability. By contrast, in the scenario which Tešić et al., (2020) predicted would reduce the production of base rates in diagnostic reasoning, and which they found did so, the causes are not interchangeable – they have names ('Tom' and 'Michael'), and it seems unlikely that the base rates given for the probability that these two people they will do as asked, and bring red wine to the party, could have been produced on the basis of a large number of trials. Such probabilities will have been produced on the basis of priors, updated whenever Tom or Michael exhibit reliability and conscientiousness, or the reverse. This is a scenario that inherently tends to a subjective interpretation of probability, and one where participants may be less afraid of not carrying out a statistical task, with a scientific flavour, in line with experimenter expectations. 'Probability' is a term with both an everyday meaning and a technical meaning (or meanings). It seems likely that many non-academic reasoners are aware of the suggestion often made that misuse and misunderstanding of probabilities is widespread. As noted above, Tešić et al., (2020) do not list the explanations given by participants for their judgements, but it seems many of them said that the 'probability' of a cause, given to them by the experimenters, was not affected by the effect. Perhaps they would have answered differently if they had been asked for the 'likelihood', or if the participants were asked not to say what they thought was the probability, but what an average person would think the probability was. These questions of wording are what I will turn to next, looking in particular at the present research. This research was carried out to see if the striking discrepancy found in Ali et al., (2011) could be

convincingly replicated, and so the wording of the scenarios in that paper were largely carried forward, filtered through pre-testing in Hall et al., (2016), to the present research. The scenarios used here are not identical to those used in the Ali et al., (2011) and Hall et al., (2016), but they match them closely, particularly in the choice of verb tenses, and lack of adverbs from which probabilities can be inferred, such as ‘often’, ‘usually’.

Thus, wording such as

If he oversleeps, then he is late. If his car breaks down, he is late.

(an example from the present research) suggests that the symbols in logical propositions (if p1 then q, if p2 then q) have been replaced with seemingly generic clauses, rather than aiming for sentences which are likely to be found in ordinary discourse. However, deliberately attempting to use non-specific language may have unexpected results due to language pragmatics: for example, ‘or’ conveys less information than ‘and’, but pragmatically, it leads to an inference about what the user knows, i.e., not enough to use ‘and’. Furthermore, this direction, from general logical statements to instances of natural language, assumes that natural language statements somehow come second in reasoning to logic or observed frequencies. Pearl (2000, p. 252-253) suggests that “the bulk of our causal knowledge” comes as “linguistic advice”, i.e., natural language statements.

Language effects can be subtle. Although hierarchical modelling techniques offer some hope of reducing the effects of unwanted differences between scenarios (Clark, 1973), they are no help when effects are not scenario-specific. A reviewer of a draft of this thesis pointed out an unintended weakness of the way, in all the experiments in the present research, responses were elicited in the change mode. Participants were given an option of 'more likely' as an answer choice - could this not be interpreted as 'more likely than not', rather than the intended 'more likely than before you were told this extra information'. Similarly, for Tešić et al., (2020),

asking 'does the probability... change... now' where two responses indicate a change (with the word 'change' in a bold font), one does not, may perhaps encourage participants familiar with multiple choice questions to choose one of the change responses, where 'do you now think it is less, equally or more likely' may be more neutral. Again, although the questions in Hall et al., (2016), which closely resemble those in the Ali et al., (2011) and the present research, were subjected to a pre-test, the questions tested were much of a muchness, particularly with respect to much use of the simple present tense, and so the pre-testing may not have been able to reveal unintended language effects.

Here I will briefly go beyond the even briefer discussion in the introduction of the way in which English verb tense use in conditionals carries information orthogonal to past / present / future, and after that suggest that Tešić et al., (2020), in their three scenarios, have distinguished two (the coins and the balls scenarios), by the choice of tenses, so as to lead participants to see these scenarios representing fixed, deterministic law-like relationships (which the scenario content will also tend to do).

Natural-language conditionals in English

Here I repeat the four conditional sentences given above in the introduction:

0] If it rains, the grass gets wet.

1] If it rains tomorrow, I'll be surprised.

2] If it snowed tomorrow, I'd be surprised.

3] If it had snowed yesterday, I'd have been surprised.

The numbering of these sentences corresponds to a terminology reasonably common in English grammar textbooks. Here are some example explanations of the meaning of these tense combinations taken from grammar textbooks.

The zero conditional:

if + present simple, then present simple

"We use the zero conditional to talk about events or situations that can occur at any time, and often occur more than once, and their results" / *"If* can be replaced by *when* in this type of conditional sentence" / "We also use the zero conditional to talk about actions which always have the same result:" Foley and Hall (2003, p. 120):

"These sentences describe what always happens in certain circumstances e.g., scientific facts." (Vince, 2008, p.70).

The first conditional:

If + present simple, then will/won't + infinitive

"We use the first conditional to describe possible future events or situations and their results:" (Foley & Hall, 2003, p. 121)

"Real conditions (conditional 1)

if X happens, Y will happen" / "These sentences describe what the speaker thinks will possibly happen as a consequence of a real situation." (Vince, 2008, p.70)

The second conditional:

If past, would + infinitive (a "conditional" tense)

"This is also known as the unlikely or improbable conditional. The second conditional has two main meanings. 1 It can describe an improbable future event or situation. The condition is unlikely to be fulfilled because the future event is unlikely to happen" / "2 It can also describe a hypothetical current situation or event, i.e., one which is contrary to known facts. It is therefore impossible to fulfil the condition" (Foley & Hall, 2003, p. 122)

"Unreal conditions (conditional 2)

if X happened, Y would happen

These sentences describe what the speaker thinks would happen in an imaginary situation." (Vince, 2008, p.70)

The third conditional:

If past tense, would + perfect infinitive

"This is also known as the past or impossible conditional. The third conditional describes a hypothetical situation or event in the past. The past situation or event is contrary to known facts, i.e., it is an unreal or impossible situation" (Foley & Hall, 2003, p. 123)

"Impossible or past conditions (conditional 3)

if X had happened, y would have happened

These sentences describe what the speaker thinks would have happened as a consequence of a situation which is in the past, so is impossible to change." (Vince, 2008, p.74).

(These four types are the basic types: both these texts describe other tense combinations, particularly for counterfactuals.)

We have seen that Tešić et al., (2020) found a difference in reasoning with pairs of common effect conditionals, such that the scenario they introduced ('red wine') did, as predicted, lead to more normative reasoning, with a reduced tendency for participants to give base rate probabilities when asked for conditional probabilities. Above I suggested that, rather than being evidence of a reduction in a 'propensity interpretation', the first two scenarios suggested a frequentist interpretation to participants, and a misunderstanding by participants of what the experimenters meant by the term 'probability', a word with both a common sense and

a technical meaning, such that they believed ‘probability’ was being used in the sense of ‘base rate’. The lack of a common sense meaning of ‘base rate’ in relation to the reliability of named human individuals made this interpretation of the experimenters’ intentions less acceptable in the third (‘red wine’) scenario.

Now we can consider the wording that Tešić et al., (2020) used in their materials.

Tešić et al., (2020) gave their participants the common-effect conditionals thus:

(for the ‘coins’ scenario)

If **both** coins land Tails, the light bulb **does not turn on**.

(for the ‘rubber balls’ scenario)

If **both** selected balls are **rubber** balls, the light bulb **will not turn on**.

(for the ‘red wine’ scenario)

Helen **will not** drink red wine if **Tom** and **Michael** *both* **do not bring** red wine to the party.

Here we can see that the two scenarios which Tešić et al., (2020) proposed would encourage the ‘propensity interpretation’ use zero conditionals (‘scientific facts’, Vince, 2008), while the scenario which they proposed would reduce the ‘propensity interpretation’ uses a first conditional / conditional (‘possible future events’, Foley & Hall, 2003). This difference is equally in line with Tešić et al., (2020)’s explanation, and also that suggested above, that the difference is between frequentist and subjective interpretations, and an associated misunderstanding by participants of what was intended by the term ‘probability’.

(A further difference may exist between the scenarios, inasmuch as, for the ‘red wine’ scenario, the two causes (Tom brings red wine / Michael brings red wine) seem more like ‘enablers’ than causes – would we call a petrol station attendant, who fills a car’s petrol tank, a ‘cause’ of the

car's movement? The cause of Helen's drinking red wine is her liking for it, and we can hardly believe the claim in the materials that she will *only* drink red wine if Tom or Michael bring it – if she finds a bottle tucked away in a cupboard, why wouldn't she drink it? However, the importance or not of this difference for participants is not clear.)

The types of scenario used (coin / balls vs. individual humans at a specific party) and also the syntax of the conditionals used (zero conditionals vs. first conditionals) both seem to make a objective, frequentist, understanding more salient to participants.

The hypothesis proposed by Tešić et al., (2020) is that participants have used, for the first two scenarios, a propensity interpretation, and that the specific interpretation that they have chosen is not among the 'set of proposals' of this type that have been put forward before, but is rather a radical approach, that the status of an effect is not an indicator of the status of a cause, or in non-causal terms, that correlation is one-way. If reasoners believe this, and take this understanding into the real world, away from psychology experiments, they will lead a confusing and dangerous life, speeding through red-lights, for example, seeing the speedometer indicating 100 m.p.h., and wishing they had some idea of how fast they might be going. They would also be unable to understand detective stories.

A manipulation that might shed light the findings of Tešić et al., (2020) would be to change the conditional types used. Tešić et al. used conditionals congruent with their scenarios: for those two which have a frequentist flavour, they presented the relationship with zero conditionals. For the scenario which does not call to mind a frequentist interpretation, they used a first conditional. Using these scenarios with the type of conditionals reversed might affect the participants' responses. Also, presenting the scenarios with questions asking what a third-person might 'now believe' might reduce experimenter demand, which seems not to be predicted by Tešić et al., (2020)'s. propensity interpretation hypothesis. Experimenter demand

might be reduced by allowing participants to use their own unconditional probabilities, elicited perhaps by counterbalancing a particular scenario type with collecting unconditional probabilities for an alternative scenario across the participants. (Such scenarios would not include coins with a 20% chance of coming up heads.)

The difference in conditional type, unremarked upon in Tešić et al., (2020), suggests that, in that research, and even more in the present replication of Ali et al., (2010) and Ali et al., (2011), not enough attention was given to the very rich meanings, semantic and pragmatic, conveyed by natural language conditionals. While the present research used conditionals that aimed for a 'neutral' valence, and further attempted to reduce the importance of language specific by the use of multi-level modelling, this may be a false hope. If Pearl (2000) is correct in the primacy of conditional sentences over probabilistic correlations, future research should aim to make an understanding of the way reasoners actually perceive the semantics and pragmatics of natural language conditionals. In particular, simply using zero conditionals as if they are neutral choice should be avoided. This was not possible in the present research, inasmuch as a central aim was to replicate Ali et al., (2010) and Ali et al., (2011).

13.4 Summary

The research reported here has brought clarity to the response mode discrepancy, first reported in Ali et al., 2011. The discrepancy appeared in two experiments, 1A and 1B, which were a translation to the medium of the internet, with non-student participants, of the experiments in Ali et al., 2011. The discrepancy survived a change to the way probability was elicited, from inputting an integer value, to indicating a position on a linear scale. Finally, an attempt was made in experiment 4 to change the materials so as to reduce or eliminate the discrepancy, to no avail.

This clarity is balanced by continuing obscurity as to the reason or reasons for the discrepancy. The failure of experiment 4 to achieve its aim must be seen as supporting a connection between the response mode discrepancy and the greater memory load imposed by the change mode. However, no mechanism for this connection has so far received support.

This curious anomaly deserves further research. The failure in the present research to successfully collect as intended for most experiments the conditional and conjunctive probabilities needed to calculate phi and the JPT is regrettable. This might be remedied in a future study. Replication seems no longer to be an issue – the response mode discrepancy is now a robust phenomenon. The speculative account given in the sections above suggests various manipulations with the aim of altering this stubbornly persistent effect. If insight were to be produced on the link between the normative and successful account of causal reasoning in the last decades, CBN theory, and the way people actually carry out causal conditional reasoning, a fundamental task of everyday life, that insight would be valuable.

References

Adams, E. W. (1996). *A primer of probability logic*.

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403–418.

Ali, N., Schlottmann, A., Shaw, A., Chater, N., & Oaksford, M. (2010). Causal discounting and conditional reasoning in children. In *Cognition and conditionals: Probability and logic in human thought* (pp. 117–134).

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thomas Aquinas, T., Treatise on Happiness, trans. by John A Oesterle (Notre Dame: Univ. of Notre Dame Press, 1983).

Bach, K. (1999). The myth of conventional implicature. *Linguistics and Philosophy*, 327–366.

Baratgin, J., Douven, I., Evans, J., Oaksford, M., Over, D., & Politzer, G. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences*, *19*(10), 547–548.

Baratgin, J., Over, D. E., & Politzer, G. (2013). Uncertainty and the de Finetti tables. *Thinking & Reasoning*, *19*(3–4), 308–328.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.

Barwise, J. & Perry, J., (1983). *Situations and attitudes*. MIT Press Cambridge, MA.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford University Press.

Birner, B. J. (2012). *Introduction to pragmatics* (Vol. 38). John Wiley & Sons.

Bloom, A. (1981). The Linguistic Shaping of Thought: A Study in the Impact of Language on Thinking in China and the West. Hillsdale, NJ: Erlbaum Associates.

Bryant, P., & Nunes, T. (2012). Children's understanding of probability: A literature review. *London: Nuffield Foundation*.

Bürkner P (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, 80(1), 1–28.

Bürkner P (2018). "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal*, 10(1), 395–411.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61–83.

Carpenter, B., Gelman, G., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A., 2017. Stan: A Probabilistic Programming Language. Journal of Statistical Software 76(1).

Chaigneau, S. E., Barsalou, L. W., & Sloman, S. A. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General*, *133*(4), 601.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Cheng, P. W., & Lu, H. (2017). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing. in *The Oxford Handbook of Causal Reasoning*, p. 65.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Comrie, B. (1986). Conditionals: A typology. *On Conditionals*, *77*, 99.

Cruz de Echeverria Loebell, N. (2018). *On the role of deduction in reasoning from uncertain premises* [PhD Thesis]. PSL Research University; Birkbeck college (London).

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*(3), 274–282.

Davis, Z., & Rehder, B. (2017). The Causal Sampler: A Sampling Approach to Causal Representation, Reasoning, and Learning. *CogSci*.

Davis, Z., & Rehder, B. (2020). *A Process Model of Causal Reasoning*.

Cruz, N., (2018). *On the role of deduction in reasoning from uncertain premises* [PhD Thesis]. PSL Research University; Birkbeck College.

De Freitas, J., DeScioli, P., Nemirow, J., Massenkoff, M., & Pinker, S. (2017). Kill or die: Moral judgment alters linguistic coding of causality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1173.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*(1), e15954.

Derringer, C., & Rottman, B. M. (2018a). Comparing Mediation Inferences and Explaining Away Inferences on Three Variable Causal Structures. *CogSci*.

Derringer, C., & Rottman, B. M. (2018b). How people learn about causal influence when there are many possible causes: A model based on informative transitions. *Cognitive Psychology*, *102*, 41–71.

Douven, I. (2017). How to account for the oddness of missing-link conditionals. *Synthese*, *194*(5), 1541–1554.

Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology*, *101*, 50–81.

Douven, I., Elqayam, S., Singmann, H., & Wijnbergen-Huitink, J. van. (2019). Conditionals and inferential connections: Toward a new semantics. *Thinking & Reasoning*, 1–41.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Edgington, D. (2011). Causation First: Why Causation is Prior to Counterfactuals. *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*, 230.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking & Reasoning*, *19*(3–4), 249–265.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*, 31–53.

Evans, J. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, *18*(1), 5–31.

Evans, J., & Over, D. E. (2004). *Oxford cognitive science series. If*. Oxford University Press. https://doi. org/10.1093/acprof: oso/9780198525134.001.

Evans, J., Over, D. E., & Handley, S. J. (2005). *Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002).*

Feng, G., & Yi, L. (2006). What if Chinese had linguistic markers for counterfactual conditionals? Language and thought revisited. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*(28).

Fetzer, J. H. (2012). *Scientific knowledge: Causation, explanation, and corroboration* (Vol. 69). Springer Science & Business Media.

Finetti, B. de. (1974). *Theory of probability: A critical introductory treatment*.

Foley, M., & Hall, D. (2003). Advanced Learners' Grammar" Pearson Education Limited. *London*.

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gillies, D. (2012). *Philosophical theories of probability*. Routledge.

Gillies, D. (2016). The propensity interpretation. *The Oxford Handbook of Probability and Philosophy. Oxford: Oxford Handbooks*.

Glymour, M. M., & Greenland, S. (2008). Causal diagrams. *Modern Epidemiology*, *3*, 183–209.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Hacking, I., & Ian, H. (2001). *An introduction to probability and inductive logic*. Cambridge university press.

Hagmayer, Y. (2016). Causal Bayes nets as psychological theories of causal reasoning: Evidence from psychological research. *Synthese*, *193*(4), 1107–1126.

Hagmayer, Y., & Waldmann, M. R. (2007). Inferences about unobserved causes in human contingency learning. *Quarterly Journal of Experimental Psychology*, *60*(3), 330–355.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, 214–219.

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).

Hattori, M. (2016). Probabilistic representation in syllogistic reasoning: A theory to integrate mental models and heuristics. *Cognition*, *157*, 296–320.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science*, *31*(5), 765–814.

Hoerl, C., McCormack, T., & Beck, S. R. (2011). *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology*. Oxford University Press.

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

Iatridou, S. (2000). The grammatical ingredients of counterfactuality. *Linguistic Inquiry*, *31*(2), 231–270.

Jackson, F. (1990). Classifying conditionals. *Analysis*, *50*(2), 134-147.

Janssen, E. M., Velinga, S. B., De Neys, W., & van Gog, T. (2021). Recognizing biased reasoning: Conflict detection during decision-making and decision-evaluation. *Acta Psychologica*, *217*, 103322.

Jeffrey, R. C. (1990). *The logic of decision*. University of Chicago Press.

Jiang, Y. (2019). Ways for expressing counterfactual conditionals in Mandarin Chinese. *Linguistic Vanguard*.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Johnson-Laird, P. N. (1995). Mental models, deductive reasoning, and the brain. *The Cognitive Neurosciences*, *65*, 999–1008.

Johnson-Laird, P. N. (2013). Mental models and cognitive change. *Journal of Cognitive Psychology*, *25*(2), 131–138.

Johnson-Laird, P. N., & Byrne, R. M. (1991). *Deduction.* Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*(4), 201–214.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Karawani, H. (2014). *The real, the fake, and the fake fake: In counterfactual conditionals, crosslinguistically*. Netherlands Graduate School of Linguistics.

Kass, R. E., & Rafferty, A. (1995). Bayes ratios. *Journal of American Statistical Association*, *90*, 773–795.

Korb, K. B., & Nicholson, A. E. (2011). Bayesian Artificial Intelligence, ser. In *Chapman & Hall/CRC Computer Science & Data Analysis*. CRC Press Boca Raton, FL.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Mit Press.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives* (Vol. 36). Oxford University Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*.

Krzyżanowska, K. (2019). What is wrong with false-link conditionals? *Linguistics Vanguard*, *5*(s3).

Krzyżanowska, K., & Douven, I. (2018). Missing-link conditionals: Pragmatically infelicitous or semantically defective? *Intercultural Pragmatics*, *15*(2), 191–211.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, *14*(4), 389–433.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. *Causal Learning: Psychology, Philosophy, and Computation*, 154–172.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. In *Ifs* (pp. 129–147). Springer.

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

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Manktelow, K., (2012). *Thinking and reasoning: An introduction to the psychology of reason, judgment and decision making*. Psychology Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information. San Francisco: WH Freeman, 1982*.

Mayrhofer, R., Goodman, N. D., Waldmann, M. R., & Tenenbaum, J. B. (2008). Structured correlation from the causal background. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*(30).

Mayrhofer, R., Hagmayer, Y., & Waldmann, M. (2010). Agents and causes: A Bayesian error attribution model of causal reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32).

Mayrhofer, R., & Waldmann, M. R. (2014). Indicators of causal agency in physical interactions: The role of the prior context. *Cognition*, *132*(3), 485–490.

McElreath, R. (2015). *Statistical rethinking: Texts in statistical science*. CRC Press Boca Raton, FL.

McGrayne, S. B. (2011). *The theory that would not die*. Yale University Press.

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227–237.

Miller, D. W. (1994). Critical rationalism: A restatement and defence. Open Court, Chicago and La Salle. *Br J Philos Sci*, *46*, 610–616.

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.

Neapolitan, R. E. (2004). *Learning bayesian networks* (Vol. 38). Pearson Prentice Hall Upper Saddle River, NJ.

Nickerson, R. (2015). *Conditional Reasoning: The Unruly Syntactics, Semantics, Thematics, and Pragmatics of" if"*. Oxford University Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, *10*(2), 289–318.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*(3–4), 346–379.

Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. *Oxford Library of Psychology. The Oxford Handbook of Causal Reasoning*, 327–346.

Oaksford, M., & Chater, N. (2020). *Integrating Causal Bayes Nets and inferentialism in conditional inference*.

Oaksford, M., & Hall, S. (2016). On the source of human irrationality. *Trends in Cognitive Sciences*, *20*(5), 336–344.

Over, D. E. (2017). Causation and the probability of causal conditionals. *The Oxford*.

Over, D. E., & Cruz, N. (2017). Probabilistic accounts of conditional reasoning. In *International Handbook of Thinking and Reasoning* (pp. 434–450). Routledge.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186–216.

Pearl, J. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, *19*.

Pearl, J. (2001). Bayesianism and causality, or, why I am only a half-Bayesian. In *Foundations of bayesianism* (pp. 19–36). Springer.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books.

Pilditch, T. D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological Science*, *30*(2), 250–260.

Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking & Reasoning*, *16*(3), 172–197.

Popper, K. (1957). The Propensity Interpretation of the Calculus of Probability, and theQuantum Theory'. *Observation and Interpretation, ButterworthsScientific Publications, London*.

Ramsey, F. P. (1931/1990). The foundations of mathematics and other logical essays. London: Routledge and Kegan Paul.

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

Rehder, B. (2015). The role of functional form in causal-based categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(3), 670.

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.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Mit Press.

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

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.

Sanford, D. (2011). *If P, then Q: Conditionals and the Foundations of Reasoning*. Routledge.

Schroyens, W. J., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking & Reasoning*, *7*(2), 121–172.

Skovgaard-Olsen, N. (2016). Motivating the relevance approach to conditionals. *Mind & Language*, *31*(5), 555–579.

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., & Klauer, K. C. (2019). Cancellation, negation, and rejection. *Cognitive Psychology*, *108*, 42–71.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, *150*, 26–36.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Discovery algorithms for causally sufficient structures. In *Causation, prediction, and search* (pp. 103–162). Springer.

Stalnaker, R. C. (1968). A theory of conditionals. In N . Rescher (ed), Studies in logical theory (pp. 98–112). American Philosophical Quarterly Monograph Series, 2. Oxford: Blackwell.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. *Reasoning: Representation and Process in Children and Adults*, 55–79.

Tešić, M., Liefgreen, A., & Lagnado, D. (2020). The propensity interpretation of probability and diagnostic split in explaining away. *Cognitive Psychology*, *121*, 101293.

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. *R Package Version*, *2*(0), 1003.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Verma, T., & Pearl, J. (1988). *Influence diagrams and d-separation*. UCLA, Computer Science Department.

Vince, M. (2008). *Macmillan English grammar in context. Advanced*. Macmillan Education.

von Fintel, K. (2011). 59. Conditionals. In *Volume 2* (pp. 1515–1538). De Gruyter Mouton.

von Sydow, M., Hagmayer, Y., Meder, B., & Waldman, M. R. (2010). How causal reasoning can bias empirical evidence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32).

Walsh, C. R., & Sloman, S. A. (2007). Updating beliefs with causal models: Violations of screening off. *MA Gluck, JR Anderson & SM Kosslyn, A Festschrift for Gordon H. Bower*, 345–358.

Walsh, C. R., & Sloman, S. A. (2004). Revising causal beliefs. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*.

Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*(1), 1–48.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82.

Woltz, D. J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General*, *117*(3), 319.

Appendices

Appendix 1: conditional sentences used in Hall et al., (2016)

Experiment 1: Common-effect (CE) conditional pairs

1. If she throws the vase, then the vase breaks
   If a tennis ball hits the vase, then the vase breaks
2. If the fuse on the stereo is blown, then the stereo is off
   If the stereo is unplugged, then the stereo is off
3. If she is being fired from her job, then she gets upset
   If she is breaking up with her partner, then she gets upset
4. If the car has a mechanical fault, then the car stops
   If the car runs out of petrol, then the car stops
5. If the battery is running out, then the watch stops
   If you forget to wind the watch, then the watch stops
6. If there is an accident on the main road, then he is caught in a traffic jam
   If there are road works on the main road, then he is caught in a traffic jam
7. If a person wants to change their career, then they book a meeting to see a careers advisor
   If a person is finishing their degree, then they book a meeting with a careers advisor
8. If he sleeps in, then he is late
   If his car breaks down, then he is late
9. If she has a day off from work, she is going shopping
   If she is paid, she is going shopping
10. If she reads the newspapers, she learns more about world issues
    If she watches the news, she learns more about world issues

Experiment 1: Common-cause (EC) conditional pairs

1. If she feels dizzy, then she is hungry
   If her stomach is rumbling, then she is hungry
2. If it is warm outside, then it is sunny
   If there are shadows, then it is sunny
3. If there are bubbles, then the water is boiling
   If there is steam, then the water is boiling
4. If I am shivering, then I am cold
   If my hairs are raised, then I am cold
5. If there are puddles in the road, then it has been raining
   If my clothes are wet, then it has been raining

6. If the food is piping hot, then the food is cooked
   If the food is golden brown, then the food is cooked
7. If your colleague has time off work, then he is still ill
   If your colleague visits the Dr, then he is still ill

Experiment 2: Common-effect (CE) conditional pairs

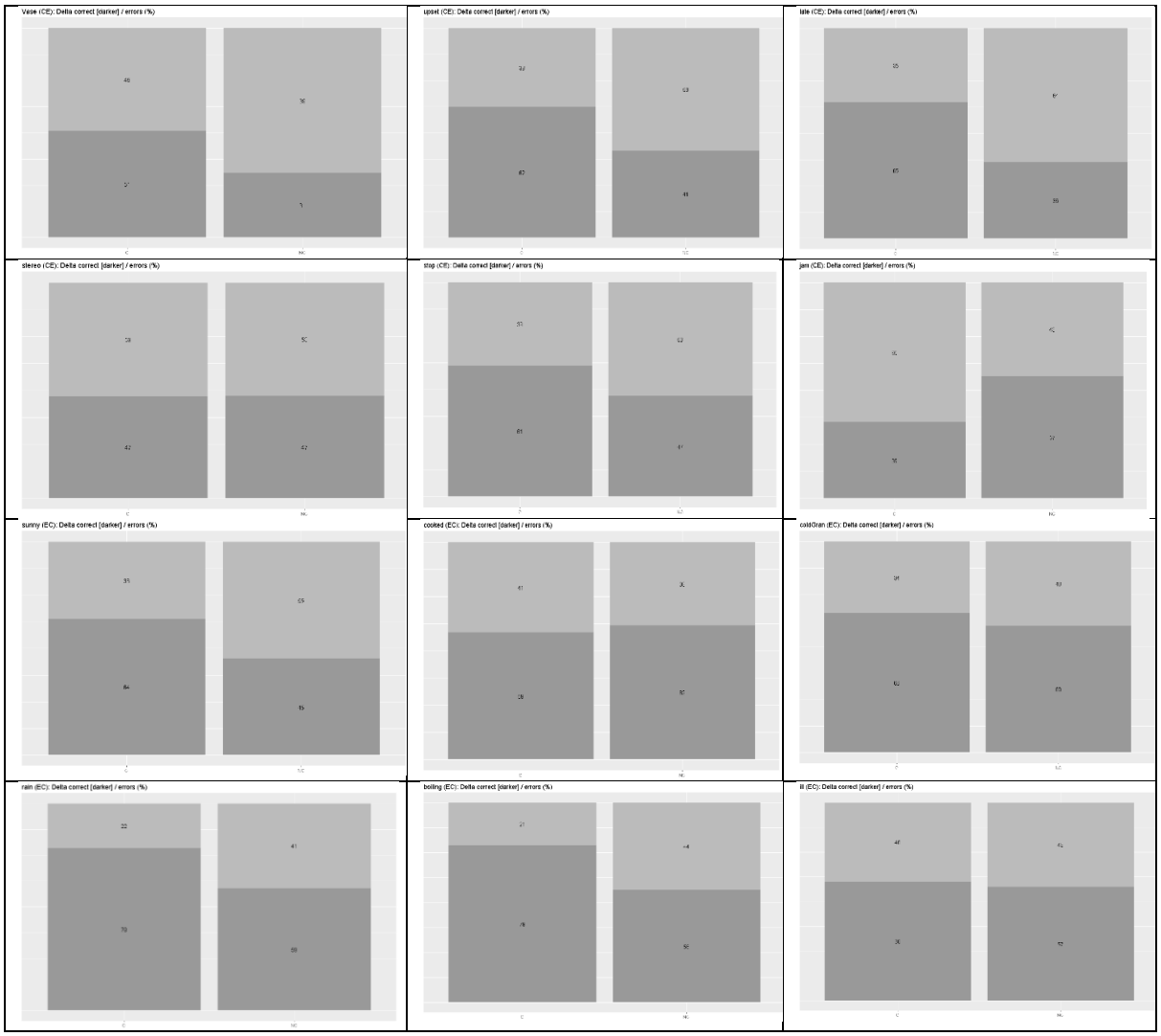
1. If she needs a break from work, then she books a cruise
   If she wants to travel abroad, then she books a cruise
2. If she shares her toys, then the teacher gives her a good report
   If she speaks politely, then the teacher gives her a good report
3. If he is tactical, then he wins the fight
   If he is determined, then he wins the fight
4. If the plant is watered often, then it grows well
   If the plant receives light, then it grows well
5. If her room is clean, then she is allowed to go out
   If her homework is complete, then she is allowed to go out
6. If he has valid insurance, then he drives the car
   If the car has a valid MOT, then he drives the car

Experiment 2: Common-cause (EC) conditional pairs

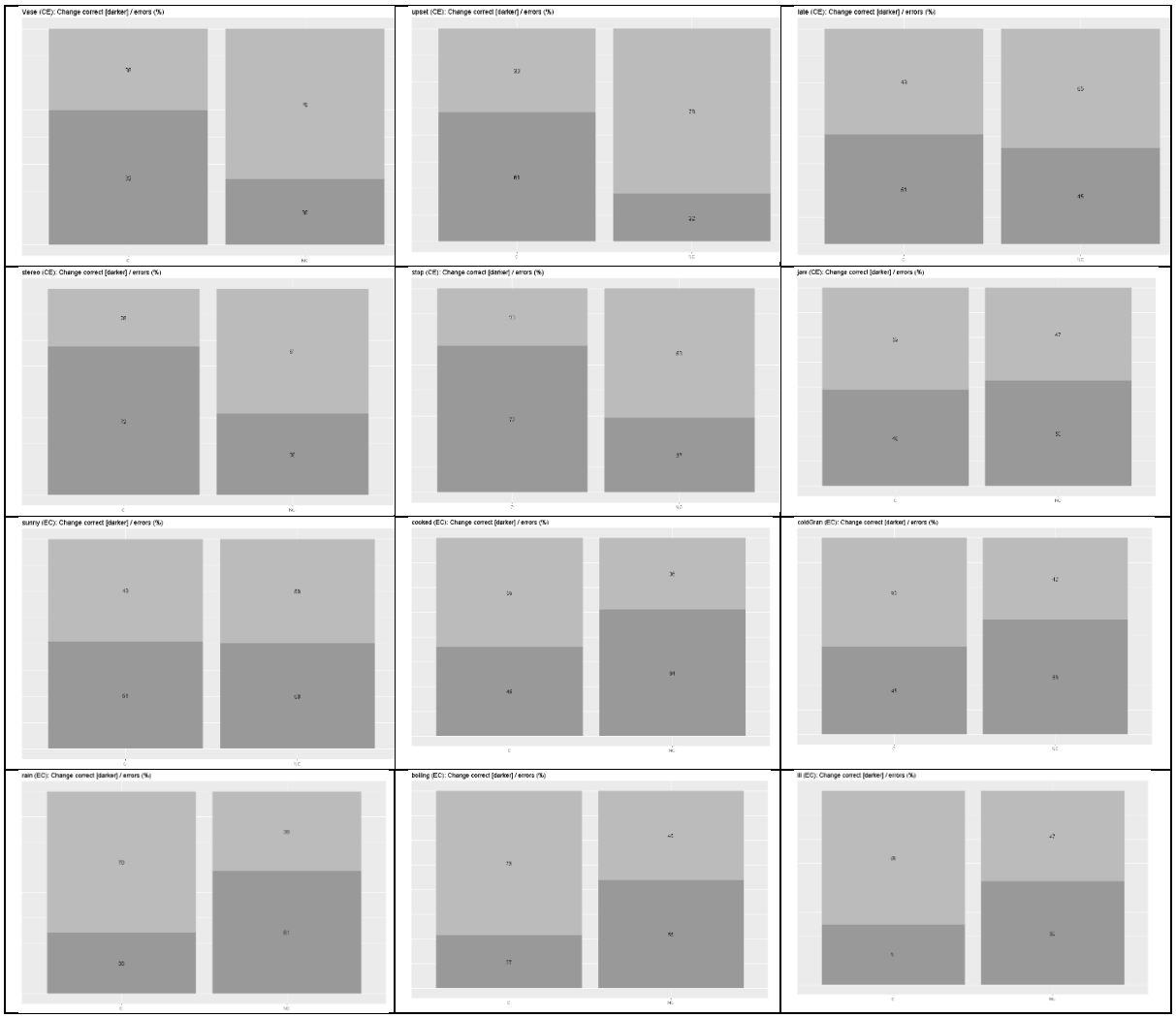
1. If the participant wears glasses, then their vision is poor
   If the participant wears contact lenses, then their vision is poor
2. If he goes to the local shop, then the milk has run out
   If he has a black coffee, then the milk has run out
3. If she is ordering coffee, then she fancies a hot drink
   If she is ordering tea, then she fancies a hot drink
4. If the food is undercooked, then the oven is faulty
   If the food is burnt, then the oven is faulty
5. If the occupier is looking to buy a new property, then she wants to move
   If the occupier is looking to rent a new property, then she wants to move
6. If she is renting DVDs for a movie night, then she is in the mood to watch a film
   If she is going to the cinema, then she is in the mood to watch a film

Appendix 2: graphs showing results for experiments 1A to 4, separated by scenario (scenario names match those in appendix 4)

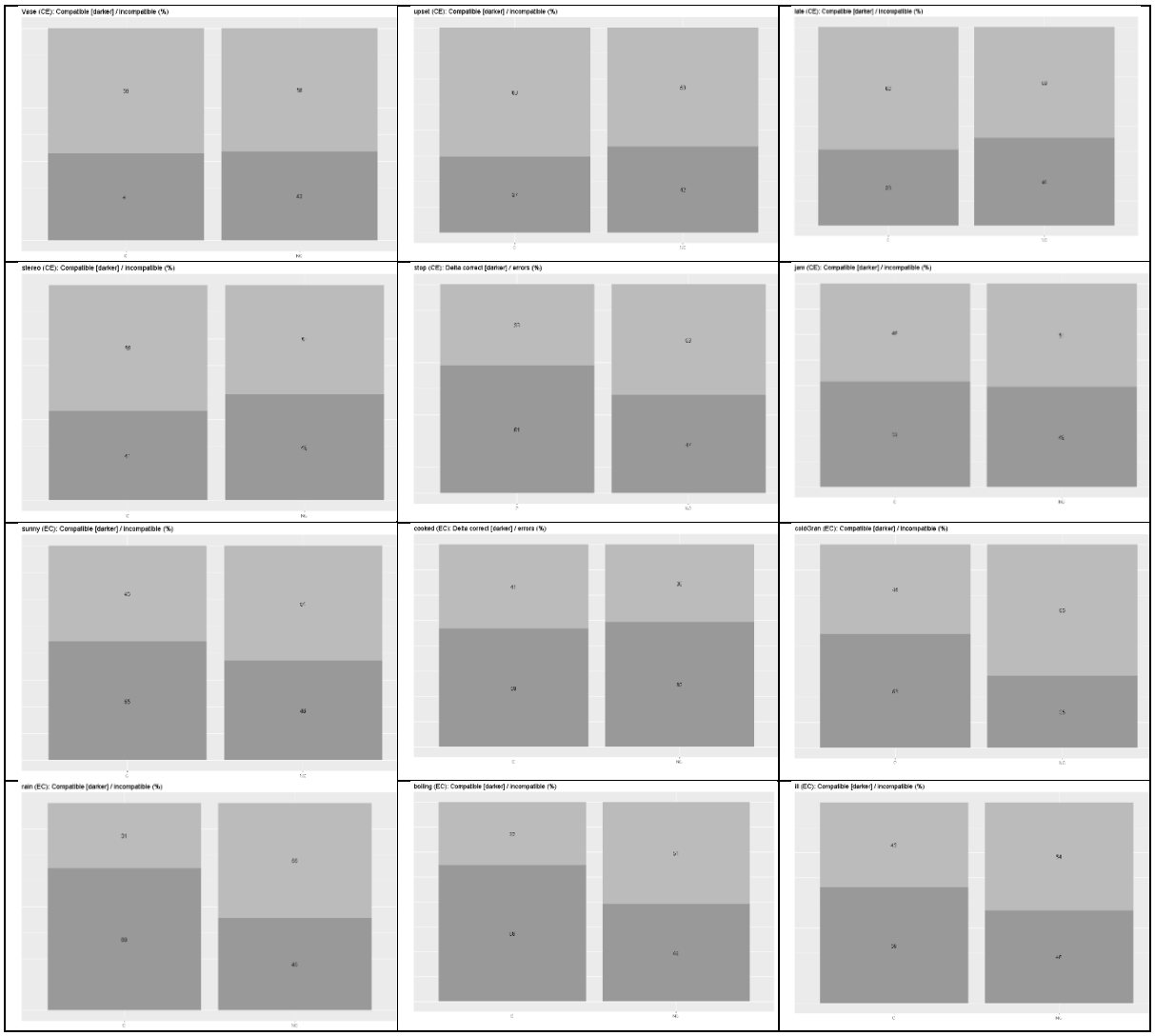
Delta errors:



Change errors:



Compatibility of responses between modes:



Appendix 3: the results in section 10.1 using an 89% HDI in place of a 95% HDI

The text in section 10.1.3, relating to each experiment, would read as follows:

Following the model comparison carried out above, for the change mode, the models examined did not include an interaction term. For the delta mode, the full models, that is to say, including an interaction of consequent and causal direction, were examined.

For the change mode, across all seven experiments, and for all four experimental conditions, none of the 89% HDI fell within the ROPE. Thus, by this metric, the CE-C and EC-NC change mode responses were normative in all experiments, since they were not equivalent to zero, and the CE-NC and EC-C change mode responses were non-normative in all experiments, since zero change is normative for these conditions.

For the delta mode, in the CE-C condition, for all seven experiments, the HDI fell completely outside the ROPE, as is normative for this condition where discounting is predicted. For the condition where augmenting was expected, EC-NC, the HDI fell completely outside the rope for six experiments: 1A, 1B, 2A, 3A, 3B, and 4. For experiment 2B, 54.28% of the HDI fell within the ROPE.

For the two conditions where neither discounting or augmenting was normative, and for which the ROPE delimited normative responses, the results were mixed in the delta mode. For the CE-NC condition, the HDI fell completely outside the HDI for experiments 1A, 2A. For experiment 1B, 1.60% of the 89% HDI fell within the ROPE, for 2B, 9.38%, for 3A, 29.94%, for 3B, 1.18%, and for 4, 13.31%. For the EC-C condition, the HDI fell completely inside the ROPE for experiments 1B, 2A, 2B, 3A and 3B, in line with normative responses. For experiments 1A, 3A, and 4, most of the HDI fell within the rope: 1A: 87.53%, and 4: 91.29%.

The text in section 10.1.5, relating to the combined data set, would read as follows:

As described above, a ROPE was set for each condition, and how much of the HDI fell within that ROPE was calculated. The differences from the analyses given above were as follows. Firstly, for both modes, the models examined were without an interaction term, in line with the model comparison. Thus the model for the delta mode was different from those used when examining each experiment separately. Secondly, the delta scores for the experiments 1A and 1B were adjusted to match the scale of the subsequent experiments while collating the individual data sets. This meant that a ROPE was chosen as described above for experiments 2, 3, and 4 in the delta mode.

For the change mode, for all four experimental conditions, none of the 89% HDI fell within the ROPE. Thus, as for the individual analyses, by this metric, the CE-C and EC-NC change mode responses were normative in all experiments, since they were not equivalent to zero, and the CE-NC and EC-C change mode responses were non-normative in all experiments, since zero change is normative for these conditions.

For the delta mode, in the CE-C condition, the HDI fell completely outside the ROPE, as is normative for this condition where discounting is predicted. For the EC-NC condition, the HDI also fell completely outside the rope. This is also normative for this condition, where augmenting is expected.

Once more, results were not conclusive for the two conditions where neither discounting or augmenting was normative, and for which the ROPE delimited normative responses. For the CE-NC condition, 5.62% of the 89% HDI fell within the ROPE. For the EC-C condition, 93.06% of the 89% HDI fell within the (normative) ROPE.

Appendix 4: example wording for each causal scenario

An example of the wording for each of the 12 scenarios, given here for the delta mode, with an integer response as in experiments 1A and 1B.

The sentences asserting the consequent (C version) or not (NC versions) are shown in bold here, but were not in bold in the experimental materials. For each causal direction (CE or EC), three scenarios are given in the C version, three in the NC version.

Common effect (CE) scenarios

Vase CE-C

R1 judgement

You have recently moved into a new flat. One morning you receive a package from your aunt. You open it up and see that she has given you and your friend a vase as a housewarming gift. Your friend thinks that this vase is especially hideous and tells you the following information: If she throws the vase, then the vase breaks. If a tennis ball hits the vase, then the vase breaks. **A few days later you are out at work and while on the phone to your flatmate, you hear in the background that the vase breaks.** You wonder whether she throws the vase. How likely do you think it is that she throws the vase? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

She tells you that a tennis ball hits the vase. How likely do you now think it is that she throws the vase? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Upset CE-C

R1 judgement

You are worried about your friend. When she is upset, she spends a lot of time crying. You are aware that:If she is being fired from her job, then she gets upset. If she is breaking up with her partner, then she gets upset. **On a particular day as you speak to her, she gets upset.** You wonder whether she is being fired from her job. How likely do you think it is that she is being fired from her job? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

She then tells you that she is breaking up with her partner. How likely do you now think it is that she is being fired from her job? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Late CE-C

R1 judgement

Your employee has an important meeting with a client. He must arrive on time to make a good impression. From past experience you know that:If he oversleeps, then he is late. If his car breaks down, then he is late. **That morning you receive a call from the client saying that your employee is late.** You wonder whether your employee oversleeps. How likely do you think it is that your employee oversleeps? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

You are now told by a colleague that your employee's car breaks down. How likely do you now think it is that he oversleeps? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Stereo CE-NC

R1 judgement

You and your housemate are having a house party and wish to play some CDs on your stereo. You take note of the following information:If the fuse on the stereo is blown, then the stereo is off. If the stereo is unplugged, then the stereo is off. **As your flatmate wanders over to the stereo to put in one of the CDs, you are unsure whether the stereo is off**. You wonder to yourself whether the fuse on the stereo is blown. How likely do you think it is that the fuse on the stereo is blown? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

He now calls back that the stereo is unplugged. How likely do you now think it is that the fuse on the stereo is blown? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Stop C-NC

R1 judgement

Your brother who has newly passed his test and does not have his own car, needs to make a long journey. He pleads with you to lend him your car and you reluctantly agree. Before he sets off, you hear your dad warn him of the following:If the car has a mechanical fault, then the car stops. If the car runs out of petrol, then the car stops. You wave your brother off as he embarks on the journey. **After an hour or so, you do not know whether or not the car stops, but you begin to wonder whether the car has a mechanical fault**. How likely do you think it is that the car has a mechanical fault? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

Your brother sends another text message revealing the car runs out of petrol. How likely do you now think it is that the car has a mechanical fault? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Jam C-NC

R1 judgement

You are meeting a friend in town and you know that he is planning to drive there. You know that:If there is an accident on the main road, then he is caught in a traffic jam. If there are road works on the main road, then he is caught in a traffic jam. **You arrive early so you do not know whether your friend is caught in a traffic jam.** However you still wonder whether there is an accident on the main road. How likely do you think it is that there is an accident on the main road? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

You now remember that there are road works on the main road. How likely do you now think it is that there is an accident on the main road? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Common cause (EC) scenarios

Sunny EC-C

R1 judgement

Watching a documentary about weather conditions, the narrator states that:If it is warm outside, then it is sunny. If there are shadows outside, then it is sunny. You have been inside all day when your younger brother returns from school. **He tells you it is sunny**. You are

wondering whether it is warm outside. How likely do you think it is that it is warm outside? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

He tells you that there are shadows outside. How likely do you now think it is that it is warm outside? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Cooked EC-C

R1 judgement

You are cooking dinner for some friends. You are following the recipe from a cookbook. The cookbook states:If the food is piping hot, then the food is cooked. If the food is golden brown, then the food is cooked. Whilst you are entertaining in the lounge, one of your guests checks on the food in the oven. **She returns and informs you that it is cooked.** You wonder to yourself whether the food is piping hot. How likely do you think it is that the food is piping hot? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

She now returns and tells you that the food is golden brown. How likely do you now think it is that the food is piping hot? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Cold gran EC-C

R1 judgement

One day you are out with your grandmother. She asks you to hold her jacket and tells you to remind her to wear it whenever she gets cold. She tells you:If she is shivering, then she feels cold. If the hairs on her arms are raised, then she feels cold. **Later on that day, she tells you that she feels cold.** You begin to wonder whether she is shivering. How likely do you think that it is that she is shivering? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

You now notice that the hairs on her arms are raised. How likely do you now think that it is that she is shivering? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Rain EC-NC

R1 judgement

You look out the window to see whether to take an umbrella out today. You know that:If there are puddles in the road, then it is raining. If the parked cars are wet, then it is raining. The window is misty so you can not see everything clearly. **You wonder whether there are puddles in the road.** How likely do you think it is that there are puddles in the road? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

You now notice that the parked cars are wet. How likely do you now think it is that there are puddles in the road? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Boiling EC-NC

R1 judgement

The kettle at work is transparent so you can see the water inside. You are making a cup of tea. You switch on the kettle and return to your desk to work while waiting for the water to boil. You ask your colleague who is not far from the kettle to call you over once it is boiled. You both know that:If there is steam, then the kettle is boiling. If there are bubbles, then the kettle is boiling. **A little while later, she leaves the room so you do not know whether or not the water is boiling**. However, you still wonder whether there is steam. How likely do you think it is that there is steam? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

Your colleague returns and now tells you that she can see that there are bubbles. How likely do you now think it is that there is steam? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

Ill EC-NC

R1 judgement

Your colleague has been feeling under the weather recently. Your boss tells you:If your colleague has time off work, then your colleague is ill. If your colleague visits the doctor, then your colleague is ill. **You receive a phone call from your colleague, however you do not know whether he is ill.** You wonder to yourself whether he has time off work. How likely do you think it is that your colleague has time off work? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):

R2 judgement

Your colleague now adds that he visits the doctor. How likely do you now think it is that your colleague has time off work? Enter a number from 0 (certainly not), 5 (completely uncertain) to 10 (certainly):