Novel perspectives on the causal mind

Experiments, modeling, and theory

Kolvoort, I.R.

NOVEL PERSPECTIVES ON THE CAUSAL MIND

EXPERIMENTS, MODELING, AND THEORY

IVAR R. KOLVOORT

Novel Perspectives on the Causal Mind: Experiments, Modeling, and Theory

Cover image by Ivar R. Kolvoort (made using Stable Diffusion XL v1.0)
Cover design by John Skead

ISBN 978-94-6473-224-5

Novel Perspectives on the Causal Mind: Experiments, Modeling, and Theory

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 9 oktober 2023, te 14.00 uur

door Ivar Kolvoort
geboren te Utrecht

Promotiecommissie

Promotores:	dr. L. van Maanen	Universiteit Utrecht
	prof. dr. H.L.J. van der Maas	Universiteit van Amsterdam
Copromotores:	dr. K. Schulz	Universiteit van Amsterdam
Overige leden:	prof. dr. D. Borsboom	Universiteit van Amsterdam
	prof. dr. D.W. Rietveld	Universiteit van Amsterdam
	dr. W.H. Zuidema	Universiteit van Amsterdam
	dr. N.R. Bramley	University of Edinburgh
	dr. J.M. Haaf	Universiteit van Amsterdam
	prof. dr. L.C. Verbrugge	Rijksuniversiteit Groningen

Faculteit der Maatschappij- en Gedragswetenschappen

The research for this doctoral thesis received financial assistance from the University of Amsterdam in the form of an Interdisciplinary Doctoral Agreement grant.

CONTENTS

Part 1: Experimental Studies on Probabilistic Causal Inference

Part 2: Computational Cognitive Modeling of Causal Reasoning

Part 3: Affordances and Causal Engagement

1 General Introduction

Causal cognition is a complex and multifaceted phenomenon that is fundamental to (human) cognition. It plays a crucial role in most of our daily activities, including decision making, problem solving, and learning. At its most basic level, it is about two events in the world, one of which is the consequence (the *effect*) of the other (the *cause*). This seems intuitive to us only because we are so experienced with it. Understanding causation is not as straightforward as it might seem. As most of us know, correlation does not imply causation, but what does imply causation is harder to establish (see Pearl, 2009). Hume (1748) famously noted that we do not directly perceive causation and that, instead, we have to infer its presence. When, for example, we see a billiard ball push another out of the way, we quickly judge that one ball *caused* the other to move, but this causation itself is not perceived, only the balls moving. While causal reasoning can seem to be an intellectual affair, also these type of quick judgments are part of causal cognition and shape the way we think and act.

We start with developing this intuitive ‘causal sense’ for the world at the earliest age, where we learn about the effects of our own actions on our own perceptions (Muentener & Bonawitz, 2017b). This starts very simple, we find out that crying causes our caretakers to give us attention and that letting go of an object causes it to fall. Over time this way of making sense of the world grows more nuanced and sophisticated, enabling us to achieve extraordinary feats. But it is not just for extraordinary feats that we use our capacity of engaging with the causal structures around us. As illustrated by watching balls collide, causal cognition is a basic component of our psychology. Whether we are watching billiard balls collide, baking cookies, or designing jet engines, our causal knowledge plays a crucial role. In our daily pursuits, we seek to comprehend why things happened the way they did and predict how we can improve them in future endeavors. For both explanation and prediction a sense of the causal structures at work is crucial. By leveraging causality we can explain why our cookies turned out poorly (“adding too much water caused the cookies to have a soup-like texture”) and use this understanding to improve our future (“next time I will buy cookies instead”). While causal cognition is not uniquely human, our capacity to control the world by exploiting causal relationships is what has made the human species thrive and manipulate the world to an unimaginable degree.

Given that we lean on our causal knowledge in almost any activity, understanding causal cognition is crucial for understanding the human mind and so an important goal for cognitive science and psychology alike. Furthermore, if we aim to design machines that can mimic human behavior, including intelligence, it’s essential that these machines also work with causality in a way that is similar to how humans do. Some have argued that this is indeed one of the things that is missing from current iterations of artificial intelligence (Lake et al., 2017).

The pervasive nature of causal cognition has led researchers from a wide variety of fields, including cognitive, social, comparative, and mathematical psychology, cognitive science, philosophy, computer science, statistics, logic, and linguistics, to show interest in the topic. With this thesis I aim to further our understanding of the *causal mind* and to improve our methods for doing so. So let us consider the ways in which causal cognition is studied within the psychological sciences and the ways in which I will build upon them.

1.1 WAYS TO STUDY AND UNDERSTAND THE CAUSAL MIND

Reflecting the ubiquity of causality in cognition, experimental methods to probe causal cognition abound (see Holyoak & Cheng, 2011; Sloman, 2009; Sloman & Lagnado, 2015; Waldmann, 2017b). Researchers can use various approaches to teach causal information to participants; via graphs, vignettes, syllogisms, abstract or naturalistic video clips, or by providing information numerically, through frequencies or probabilities. And of course, various combinations of these approaches are possible. Similarly, there are a variety of methods to elicit responses from participants, such as asking them to reconstruct the underlying causal structure is, intervene in a causal system, to judge the strengths of causal relationships, to estimate the probabilities of particular causal variables (e.g. particular effects or causes) occurring, and more. Moreover, due to its pervasive nature, researchers can indirectly probe causal cognition by having people provide responsibility or moral judgments, categorize objects, or by having participants explain why things happened.

In many experiments, including those presented in this thesis, researchers use a combination of statistical information and graphical representations to teach participants about causal systems and to elicit responses. Figure 1.1 provides such a graphical representation of a 3-variable causal structure, with nodes (circles) representing causal variables and lines indicating causal relationships. The experiments presented in Chapters 2 and 3 employ such a graphical representation, and participants are asked to make causal probabilistic inferences. Such inferences are typically of the form (referring to the structure in Figure 1.1): “Currently people are swimming, but you do not know whether it is hot weather. What is the probability that people are eating ice-cream?”. That is, participants are provided with some information about an instantiation (or ‘case’) of the causal network (e.g. that people are swimming), and are then asked to infer the state of another variable in the network (e.g. whether people are eating ice-cream). To arrive at a good judgment, participants must draw on their understanding of the network of causal relations.

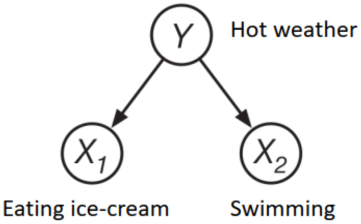


Figure 1.1 Example of a causal network, where the circles refer to causal variables and the arrows represent causal relationships, pointing from cause to effect. Here it is a 3-variable ‘common cause’ structure, where one cause (Y, hot weather) produces 2 effects (X_1, eating ice-cream; X_2, swimming).

Experimental methods such as these have allowed researchers to probe causal cognition from multiple directions. One thing common to these different experimental designs is that responses obtained from them are mostly interpreted using the theory of Causal Bayesian Networks (CBNs; Pearl, 1988, 2009; Spirtes et al., 2000). Moreover, the use of these experimental designs often goes hand in hand with the implicit assumption that human causal cognition follows CBN theory to some degree, as the way participants are presented with information (using graphs and statistical information) is derived from CBN theory. Hence, to understand the psychological literature on causal cognition, one needs to know a bit about CBN theory.

Specifically, CBNs are a mathematical and graphical formalism that offer a concise representation of causal knowledge and a formal logic that specifies how one can learn, draw inferences, imagine counterfactuals, and update causal knowledge based on interventions in a causal system. A CBN consists of a causal network structure (as in Figure 1.1) in addition to a joint probability distribution specifying the likelihoods and dependencies of the variables in the graph. For example, part of the joint probability distribution would be the probability that someone eats ice-cream while it is hot weather, which could be e.g. 80%. We would write this in mathematical notation as $P(X_1 = 1|Y = 1) = .80$. A CBN model would also specify the probability that someone eats ice-cream while it is not hot outside (e.g. $P(X_1 = 1|Y = 0) = .15$), in addition to the probabilities for any other combination of variable values.

CBN theory firstly is a normative account of how one *should* reason about causality, providing a consistent logic for reasoning with causal information. Since their development, however, psychologists and cognitive scientists have been able to use CBNs descriptively to model human behavior with remarkable success. CBN-based models have been shown to describe a variety of human behaviors related to causal learning, inference, reasoning, structure induction, categorization, and more (e.g. Ali et al., 2011; Bramley et al., 2015, 2017; Cheng, 1997; Coenen et al., 2015; Fernbach & Erb, 2013; Griffiths & Tenenbaum, 2005, 2009; Hagmayer, 2016; Hayes et al., 2014; Holyoak et al., 2010; Krynski & Tenenbaum, 2007; H. S. Lee & Holyoak, 2008; Lu et al., 2008; Meder et al., 2014; Rehder, 2014; Rehder & Burnett, 2005; Shafto et al., 2008; Sloman, 2009; Sloman & Lagnado, 2015; Steyvers et al., 2003; Waldmann & Hagmayer, 2006).

While CBNs have provided a crucial impetus and guidance to research into causal cognition, and have had much success in describing human behavior, there are multiple reasons for why it does not provide a satisfactory account of human causal cognition on its own. These reasons have to do with CBNs predictive accuracy, but also the nature of the formalism and the cognitive models derived from it. This thesis addresses and builds on these limitations in various ways.

As mentioned, CBN theory has had success in describing, i.e. predicting, human behavior. However, recent research has found multiple systematic ways in which people deviate from CBN predictions (e.g. Rehder, 2014; Rottman & Hastie, 2016). As CBN theory is a normative logic, we can use it as a benchmark for human behavior, and describe systematic deviations of human responses as 'errors' (being overly conservative for example). I experimentally investigate these errors and how they change under time pressure in Chapter 2. Another problem with CBN as a description of human behavior is that CBN, as a normative theory, predicts a *single* correct answer for causal inferences, yet previous experiments have shown that causal judgments vary considerably (e.g. Rehder, 2018; Rottman & Hastie, 2016). In Chapter 3, I present an experiment that elicits repeated responses to investigate the sources and systematicity of this variability.

Taking a broader view, it becomes apparent that CBNs have more significant limitations in helping us understand the causal mind beyond their predictive shortcomings (i.e. not capturing systematic deviations from the point prediction and variable judgments). CBNs provide a *computational level* perspective on causal reasoning (Marr, 1982), meaning they describe the computational problem an agent is solving when engaged in causal reasoning. This gives us a sense of *what* the human mind is doing, but not *how* it is doing so. This latter question is important for understanding *how* the mind works. To that end, we need *algorithmic level* theories that explain the step-by-step *processes* by which people form beliefs or make judgments. In other words, we need *process-level* models. In Chapters 4 and 5, I present research using computational cognitive modeling to differentiate between process-level models of causal reasoning, and in doing so make progress in addressing the *how* question.

Stepping back even further, we must acknowledge that (CBN-derived) cognitive models of causal reasoning abstract away from many complexities that are essential for understanding the human mind. The way we learn about and interact with the world is constrained by our bodies and shaped by sociocultural practices (Chemero, 2009; Rietveld & Kiverstein, 2014; Shapiro & Spaulding, 2021). CBNs cannot give a satisfactory account of such processes as they exclusively model statistical regularities in an isolated causal system external to the agent. Ultimately, we would want our theories to be plausible on the *implementational level* (Marr, 1982), that is, possible to be implemented in the hardware we humans use. But what is the hardware we humans use? One answer would be 'the brain', but that is too simplistic. Philosophers working on embodied cognitive science would argue that the mind needs to be understood in relation to its environment including the body (e.g. Chemero, 2009; Gallagher & Zahavi, 2008; J. Gibson, 1979; Merleau-Ponty, 1962; Noë, 2004; Rietveld & Kiverstein, 2014; Varela et al., 1991). Brains have never been separated from bodies (at least not when they are functioning as they normally do), and bodies have never been separated from their environments (I would not even know what that means). Based on these observations (and many others), theorists working on *embodied cognition* argue that cognition does not take place in the skull, but rather in the interaction between the agent and environment. For example, the writing of this thesis is a cognitive activity, and I need a lot more to do it than just a brain (some examples are: arms, hands, a neck, a computer, word processing software, statistics software, a keyboard, my supervisors, a university, and possibly my pancreas).

It seems clear to me that the interactions that form the relationship of the mind to the environment are of a nature and complexity that seem impossible to capture in a graphical formalism like CBN. This does not negate the utility of CBNs in understanding certain aspects of cognition, nor does it negate the utility of using an information-processing metaphor to understand the mind, as is done traditionally in ("dis-embodied") cognitive psychology. However, traditional cognitive psychology cannot help us understand many interesting mental phenomena, such as how our mind is shaped by our bodies, cultures, and the environment at large. While I believe that abstracting away from these complexities can be useful (I do so myself in Chapters 2 through 5), when doing so we ought to remind ourselves that we are simplifying that which we aim to understand. This is not just a theoretical argument. Research has, for instance, shown that our culture impacts how we experience causality (Bender et al., 2017; Bender & Beller, 2011; Morris et al., 1995). Therefore, in the last part of this thesis (Chapters 6 and 7) I take a radical turn and move away from the CBN formalism and traditional cognitive psychology to understand causal cognition in light of the full human-environment system.

As can be gleaned from this discussion, I believe strongly that a pluralistic and multidisciplinary approach is essential for understanding the many intricacies of the mind. Mental phenomena are hard to grasp and vastly complex, and so it seems hubristic to assume that a single theoretical framework can provide a comprehensive understanding of the mind. This rings particularly true for causal cognition due to its multi-faceted and pervasive nature. Hence, over this thesis as a whole, I aim to respect this inherent complexity and refrain from reducing mental phenomena related to causation, falsely, to a low-dimensional problem.

This multidisciplinary approach will be visible throughout this thesis by my use of multiple perspectives and methodologies. I began this research project by surveying the literature on causal cognition to identify gaps and shortcomings in the field while paying specific attention to the variety of methodologies used. This was fruitful but not easy, as the literature on causal cognition is dispersed across multiple disciplines ranging from psychology to philosophy, linguistics and logic. However, this dispersion makes for fertile ground for methodological and theoretical cross-pollination. That is, by leveraging my own interdisciplinary education and the dispersion of causal cognition research, I was able to identify and employ methods from adjacent fields to resolve questions or issues that have come up in the study of causal cognition. Examples of these are the use of experimental techniques such as time pressure (Chapter 2) and repeated measurement designs (Chapter 3), the use of simulation-based inference methods for testing computational cognitive models (Chapters 4 and 5), and using the concept of affordances (Chapter 6; J. Gibson, 1979) to understand how we experience and use causality (Chapter 7).

1.2 OVERVIEW OF CHAPTERS

The main body of this thesis is structured in three parts consisting of two chapters each. I will now provide a brief overview of the chapters, discussing the main motivations for each study and summarizing the results.

Part 1: Experimental Studies on Probabilistic Causal Inference

Part 1 presents two experimental studies on probabilistic causal inference. In these experiments I teach participants causal network information and then ask them to solve inference problems in the form of causal probabilistic queries (these are of the form illustrated in the introduction, for instance "if hot weather causes people to swim and eat ice-cream, what is the probability of someone eating ice-cream if you also know that they swam that day?").

Chapter 2 explores the effects of time pressure on reasoning errors to shed light on the mechanisms responsible for them. The reasoning errors here refer to certain deviations from the normative CBN model, namely conservatism, Markov violations, and failures to explain away (see Rottman & Hastie, 2016). This study was motivated by the fact that CBN is generally used as a benchmark and so most theories in the field have been developed particularly to explain these deviations from it. We manipulated time pressure to investigate what would happen to these reasoning errors, as time pressure has helped in other fields to better understand the underlying cognitive mechanisms (e.g. Evans et al., 2009; Forstmann et al., 2016; Furlan et al., 2016; Kocher & Sutter, 2006; Mulder et al., 2014; Rubinstein, 2007). We find that participants displayed increased conservatism under time pressure, and that this conservatism was related to participants' lack of confidence in their answers. This indicates that conservative causal inferences are likely the result of a general phenomenon related to uncertainty. Next, we did not find Markov violations and failures to explain away to be affected by time pressure. This was surprising as existing theories of causal reasoning, as far as it is possible to derive temporal predictions from them, seem to predict that Markov violations would increase under time pressure. Specifically, standard readings of sampling-based theories (as the Mutation Sampler and the Bayesian Mutation Sampler) and heuristic explanations of Markov violations would imply that they increase. One explanation might be that Markov violations result from processes so rapid that they are insensitive to time pressure. Together, the findings that time pressure impacts certain response patterns but not others, indicate that causal inferences (and errors) are not the result of a single cognitive mechanism.

Chapter 3 presents an experiment that uses multiple techniques to elicit repeated judgments from participants. While multiple researchers have noted that probabilistic causal judgments are remarkably variable (Rehder, 2018; Rottman & Hastie, 2016), this had never been explicitly studied. I considered this a missed opportunity, as the analysis (and modelling) of distributions of response data has allowed for theoretical developments in other domains of psychology (e.g. Bogacz, Wagenmakers, et al., 2010; Van Maanen et al., 2011; van Ravenzwaaij et al., 2011). Hence, I developed an experiment to investigate the variability present in people's causal judgments and to test whether such variability could be informative of underlying cognitive mechanisms.

The results, for the first time, established that the observed variability is due to both between- and within-participant variability in responses. Moreover, our analyses indicated that the within-participant variability is affected by the type of inference participants are asked to make. Importantly, this means that the variability in causal judgments, at least partly, reflects decision-making processes and not just noise. As such this study paves the way for future research to use the variability in responses to distinguish between theories of causal reasoning (as I aim to do in Chapter 5). Additionally, these results form a strong argument that theories only describing averaged behavior do not suffice for understanding causal cognition. Instead, we should also take into account (aspects of) distributions of responses when testing our models, which I do in Part 2.

Part 2: Computational Cognitive Modeling of Causal Reasoning

In Part 2 I develop a computational cognitive model of causal reasoning called the Bayesian Mutation Sampler (BMS) and test it against other models on the data from the experiments in Part 1. I start Chapter 4 by scrutinizing a recent model of causal reasoning, the Mutation Sampler (MS; Davis & Rehder, 2020). The MS is based on the idea that humans use a sampling mechanism to approximate Bayesian inference (Bonawitz, Denison, Gopnik, et al., 2014; Hertwig & Pleskac, 2010; Lieder & Griffiths, 2020; Vul et al., 2014). It proposes that instead of doing Bayesian calculations directly, we sample concrete instantiations of a causal system via a Markov chain Monte Carlo process to make causal judgments. My analysis identifies that, while the MS performs well at predicting mean judgments, it fails to account for salient features of distributions of causal judgments, such as a lack of extreme responses (i.e. responses near 0% and 100%). I argue that the particular misfits of the MS are due to the model lacking a mechanism for incorporating prior information. In addition to these misfits, I provide arguments for such a mechanism based on the fact that people have been shown to reason in a Bayesian fashion (i.e. with the use of a prior) in many other domains (e.g. Hemmer et al., 2015; Tauber et al., 2017; Welsh & Navarro, 2012; Zhu et al., 2020). For these reasons I develop a generalization of the MS, the BMS, which combines the sampling procedure of the MS with the use of a generic prior. I then test the MS and BMS on the experimental data from Chapter 2. I find that the BMS clearly outperforms the MS, in terms of predicting mean judgments as well as distributions of judgments. As it stands, the BMS is the first model that is able to account for response distributions on probabilistic causal reasoning tasks. This is not an easy feat considering the model accounts for full distributions for multiple different inferences using only two free parameters. These results suggest that the variability observed in causal judgments is due to the stochastic sampling scheme as proposed by the BMS, something I test further in Chapter 5.

In Chapter 5 I test the BMS against other candidate models to see whether they can account for the variability in causal judgements and other well-known patterns in causal judgment data from Chapter 3. In addition to the BMS, I tested the Beta Inference Model (Rottman & Hastie, 2016) and four other models I develop based on general psychological mechanisms that could produce variability in causal judgments. While there are many other descriptive theories of causal reasoning in the literature, the ones I test seem the only ones that can produce variable judgments as we observe them in experiments. I find that, overall, the BMS outperforms all other models. Both in terms of quantitative fit and in terms of accounting for qualitative patterns of interest the BMS fares best. One feature of the data that the BMS (and the other models) does not describe well are the changes in within-participant variability over different inference types as observed in Chapter 3. As the experiment in Chapter 3 was the first to establish that there is substantial within-participant variability, these empirical patterns have only been observed once and so they require further validation. There are options for the BMS to be extended to capture these patterns though, for example, by letting the amount of samples a reasoner takes vary based on the inference type (as suggested in Zhu et al., 2020). We provide more options for improving the fit at the end of the chapter. But, even before that work is done, the BMS already seems to provide the best process-level account of causal reasoning in the literature.

Part 3: Affordances and Causal Engagement

As mentioned in the introduction, I take a radical turn in Part 3 and make use of conceptual tools from philosophy in the tradition of radical embodied cognition. Chapter 6 presents a general introduction to the Skilled Intentionality Framework (SIF; van Dijk & Rietveld, 2017), a conceptual framework combining enactivism (Froese & Di Paolo, 2011; McGann et al., 2013; Myin, 2016; Noë, 2004) and ecological psychology (J. Gibson, 1979) to understand the situated and embodied mind. Central to SIF is a relational concept of *affordances* (Rietveld & Kiverstein, 2014), which refer to *possibilities for action*. Following SIF, affordances are relations between the sociomaterial environment and abilities available to an organism. For example, a cup affords grabbing (to me but not to an ant) and my computer (amongst other things) affords me writing this thesis. Using affordances in a relational fashion (i.e. relating organism and environment) allows for analyzing any type skilled behavior using affordance and for integrating the embodied and situated human mind at multiple scales. In this chapter I illustrate the possibility of using affordance-based analyses at multiple scales by discussing how affordances play a role on the level of our ecological niche, at the level of a sociocultural practice, and at the level of an individual.

Next, in Chapter 7 I develop and present an affordance-based account of causal engagement emphasizing the embodied and situated nature of causal cognition. Causal engagement, as I construe it here, underlies most of causal judgments and perceptions as they occur in daily life. At the core of my account is that causal engagement is a skill and this skill is about selectively attending to aspects in our environment that allow for effective interventions. These effective interventions are understood as relevant affordances. Which actions are effective interventions (or: which affordances are relevant) depends on the material and sociocultural environment. Construing causal engagement this way allows us to understand the variation in causal judgments between different cultures and between people part of different practices, as being due to differences in skills, practices, and culture. Interventions that are used in one practice might not be in another, and so people inhabiting the former might experience causality in different aspects of the environment. This is illustrated by a famous example from Carnap (1966) in which a policeman, a road engineer, and a psychologist visit the scene of a car crash. Carnap mentions that we should not expect these different people to judge the cause of the car crash to be the same, the policeman is likely to say the cause was the driver's speeding, while an engineer would point out the state of the road, and the psychologist the mental state of the driver. I argue this is due to these individuals being part of different practices in which they have developed their skill in causal engagement to intervene onto different aspects of the environment. The policeman intervenes on people's speeding (by writing tickets), the engineer intervenes on the road (e.g. by filling potholes), and the psychologist intervenes on mental states (by therapy).

My affordance-based account of causal engagement provides a theoretical framework for understanding how we experience causation and it has a broader scope than Chapters 2-5. Chapters 2-5 used the traditional conceptual framework of cognitive psychology, which conceives of causal cognition primarily in terms of processing of (statistical) information. However, solely using an information processing metaphor to understand the mind prohibits grasping the embodied, situated and, enacted nature of how we deal with causality in daily life. My affordance-based account is not at odds with the conceptual framework used in Chapters 2-5, but encompasses it and describes causality and its role in cognition at a more fundamental level. The view used in Chapters 2-5 focuses on our immensely impactful capacity to use statistical information and judge probabilities in the context of causal structures. However, there is more to causal cognition (and the mind) than that, and my aim in this chapter is to highlight those aspects that traditional cognitive psychology does not capture. In doing so, my account foregrounds the role of sociocultural context, skills, and concrete possibilities for action in what we experience as causal.

Finally, in Chapter 8 I present a brief general discussion of the work presented in this thesis, where I focus on the results and possibilities for future research. I end with a speculation on how to possibly integrate sampling-based accounts of cognition (of which the BMS is an implementation) and frameworks of ecological psychology and enactivism.

Part 1

Experimental Studies on Probabilistic Causal Inference

2 PROBABILISTIC CAUSAL REASONING UNDER TIME PRESSURE

Abstract

While causal reasoning is a core facet of our cognitive abilities, its time-course has not received proper attention. As the duration of reasoning might prove crucial in understanding the underlying cognitive processes, we asked participants in two experiments to make probabilistic causal inferences while manipulating time pressure. We found that participants are less accurate under time pressure, a speed-accuracy-tradeoff, and that they respond more conservatively. Surprisingly, two other persistent reasoning errors - Markov violations and failures to explain away - appeared insensitive to time pressure. These observations seem related to confidence: Conservative inferences were associated with low confidence, whereas Markov violations and failures to explain were not. These findings challenge existing theories that predict an association between time pressure and all causal reasoning errors including conservatism. Our findings suggest that these errors should not be attributed to a single cognitive mechanism and emphasize that causal judgements are the result of multiple processes.

This chapter has been adapted from: Kolvoort, I.R., Fisher, E.L., Van Rooij, R.A.M., Schulz, K., & van Maanen, L. (Under review). Probabilistic Causal Reasoning under Time Pressure. Preprint DOI: 10.31234/osf.io/ej26r

2.1 INTRODUCTION

Humans are expert causal reasoners, even though they might not be explicitly aware of it. The point that causal judgements play a role in many decisions, has been made many times before (e.g. Hagmayer & Osman, 2012; Rottman & Hastie, 2014). Nevertheless, it is a point worthy of reiterating here: Most actions are based on perception of and reasoning about causes and effects in the world. How will your colleagues react if you are late for that meeting? What are the chances of getting the flu knowing your flatmate has it? The answers to these and most similar questions depend on your beliefs about how events are causally related.

Numerous experiments have shown that causal reasoning affects categorization, category-based inferences, learning, prediction, as well as decision-making (e.g. Gerstenberg et al., 2021; Rottman & Hastie, 2014; Sloman & Lagnado, 2015; Waldmann & Hagmayer, 2013). Beliefs about causal structures are crucial in decision-making. Someone who planned to go mountaineering next week might decide to stay at a hotel to decrease the risk of catching the flu from a flatmate. However, had she believed that catching the flu is not caused by exposure to the flu virus then she would have decided to stay at home. This is just one example of how beliefs about causal relationships impact decision-making. Causality ties into most of what we do and think, which is why the topic of causality has received more and more attention from cognitive scientists in the last decades.

Causal Bayesian Networks[1] (CBN) have become the dominant framework for modelling probabilistic causal phenomena. CBNs have been used in many scientific disciplines as a normative framework to make predictions about causal phenomena (Koller & Friedman, 2009; Pearl, 2009; Spirtes et al., 2000). Besides being used as a normative framework, CBNs have also been used as psychological models of causal reasoning (Ali et al., 2011; Fernbach & Erb, 2013; Hagmayer, 2016; Hayes et al., 2014; Holyoak et al., 2010; Krynski & Tenenbaum, 2007; H. S. Lee & Holyoak, 2008; Meder et al., 2014; Oppenheimer, 2004; Rehder, 2014), of causal learning (Bramley et al., 2015; Cheng, 1997; Coenen et al., 2015; Gopnik & Schulz, 2007; Griffiths & Tenenbaum, 2005, 2009; Lu et al., 2008; Steyvers et al., 2003), and of categorization (Kemp et al., 2012; Oppenheimer et al., 2013; Rehder & Burnett, 2005; Shafto et al., 2008; Waldmann & Hagmayer, 2006). CBN is a normative theory in that, under the assumption that the structure and parametrization of a graph correspond truthfully to the world, the inferences the model allows for correspond truthfully to the world. CBN provides a reasonable description of human causal judgements (Hagmayer, 2016; Rottman & Hastie, 2014). CBN accurately predicts that people are susceptible to subtle changes in graph structure and parametrization (Ali et al., 2011; Fernbach & Erb, 2013; Rehder, 2014, 2018; Rottman & Hastie, 2014, 2016). However, human causal judgments do not appear to be fully in line with the normative model. Instead, they deviate persistently and systematically from the CBN prediction.

[1] This framework is also known as Bayes' Nets, Graphical Probabilistic Models or Causal Graphical Models.

2.1.1 Reasoning errors

There are three specific reasoning errors (i.e. deviations from CBN predictions) people are known to commit: violations of Markov independence, failures to explain away, and conservative inferences (Rottman & Hastie, 2014, 2016). These deviations from the CBN model are important as they can provide insight in the cognitive processes involved in causal judgments. In the remainder of this section and the rest of this manuscript we will restrict our focus on binary causal variables with generative causal relationships. Such a setting has been used as the standard in the literature on causal judgments and simplifies our discussion.

People have been found to systematically violate the principle of Markov independence ('Markov violations'[2]) in a variety of experimental paradigms and regardless of how they learn about a causal network (Ali et al., 2011; Kolvoort et al., 2021; Mayrhofer & Waldmann, 2015; Park & Sloman, 2014; Rehder, 2014, 2018; Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016; Sloman & Lagnado, 2015; Waldmann et al., 2008). Markov independence refers to the independence of certain events within a causal structure. For instance, with a common cause structure $X_1 \leftarrow Y \rightarrow X_2$ (Figure 2.1A) people often think that the state of X_1 is relevant in any situation when inferring X_2. However, X_1 is only relevant here when Y is unknown. When Y is known, information about X_1 does not provide additional information about X_2, as Y completely mediates the effect of X_1 on X_2. The exact same holds for a chain structure (Figure 2.1B).

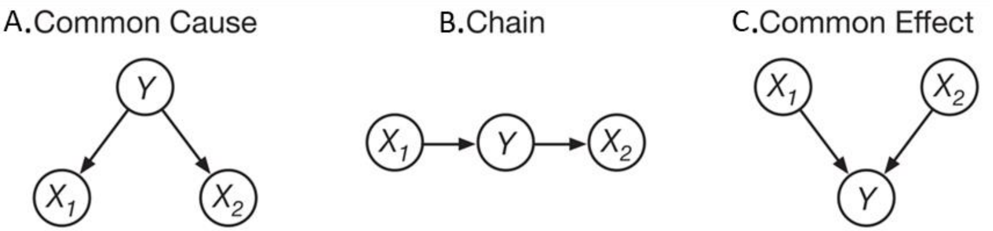


Figure 2.1 *Three-variable causal network structures. The circles represent causal variables and the arrows the causal relationships between them, pointing from cause to effect.*

In these cases we can state Markov independence formally as $P(X_i = x | Y = y, X_j = 0) = P(X_i = x | Y = y, X_j = 1) = P(X_i = x | Y = y)$, where the subscripts i and j refer to the two X variables, and the values 0 and 1 refer to a variable being absent or present respectively. Markov independence also holds in a *common effect* structure (Figure 2.1C), where the causes are independent when the effect is not known, such that $P(X_i = x | X_j = 1) = P(X_i = x | X_j = 0) = P(X_i = x)$. Instead of adhering to the principle of Markov independence, people tend to judge that $P(X_i = 1 | Y = y, X_j = 1) > P(X_i = 1 | Y = y) > P(X_i = 1 | Y = y, X_j = 0)$ in common cause and chain structures, and $P(X_i = 1 | X_j = 1) > P(X_i = 1) > P(X_i = 1 | X_j = 0)$ in the common effect structure (Ali et al., 2011; Mayrhofer & Waldmann, 2015; Park & Sloman, 2014; Rehder, 2014, 2018; Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016; Sloman &

[2] Markov violations in common cause and chain structures are also known as 'failures in screening-off', which refers to the fact that the dependence between two variables is 'screened-off' by a third variable.

Lagnado, 2015). A second reasoning error is related to the principle of *explaining away*[3]. Explaining away is involved in situations where multiple causes can independently bring about an effect and a judgment is required about the actual cause of the effect. Imagine a situation in which a friend has a headache and you know that the (only) two possible causes for this are alcohol consumption and the flu (a common effect structure, Figure 1C). Now, upon learning that your friend consumed alcohol last night it becomes less likely that they have the flu. This is because alcohol consumption 'explains away' the presence of a headache. Reversely, if you learn that your friend did not consume alcohol, it makes it more likely that they have the flu, as some other cause than alcohol consumption must have brought about the headache.

Put more generally, explaining away refers to cases in which a target cause (flu in the previous example) becomes less (more) likely after learning about the presence (absence) of another cause (alcohol consumption in the previous example, we refer to this as the non-queried cause). Referring to the common effect structure in Figure 2.1C, we can state explaining away formally as:

$$P\left(X_i = 1 \middle| Y = 1, X_j = 0\right) > P(X_i = 1|Y = 1) > P\left(X_i = 1 \middle| Y = 1, X_j = 1\right)$$

Multiple studies have found that people engage in insufficient explaining away compared to the CBN prediction or that they do not explain away at all (Fernbach & Rehder, 2013; Khemlani & Oppenheimer, 2011; Morris & Larrick, 1995; Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016; Sussman & Oppenheimer, 2011).

The third persistent reasoning error, *conservatism*, refers to a tendency of people to not give 'extreme' responses, but rather to respond somewhere near the middle of a response scale. For probability judgements this means that extreme responses near 0% or 100% are often avoided and that people prefer to respond closer to 50%. In their review of experiments on inferences from causal networks, Rottman and Hastie (2014) concluded that many inferences were conservative relative to the CBN prediction (see Baetu & Baker, 2009; Fernbach et al., 2011; Meder et al., 2008; Morris & Larrick, 1995). They found that responses are generally between 50% and the CBN prediction, which indicates that participants are not sensitive enough to the parameters of causal networks. In later work Rottman and Hastie (2016) found that judgements were particularly close to 50% when the state of one variable in the network was unknown ('ambiguity trials', such as $P(X_i = 1|Y = 1)$) or when the two other variables provided conflicted cues ('conflict trials', such as $P\left(Y = 1 \middle| X_i = 1, X_j = 0\right)$).

It is important to investigate in what situations these reasoning errors are prominent and how they come about as this can shed light on the processes underlying causal reasoning. One fruitful way to investigate these errors and accuracy in causal judgements more generally is by utilizing time pressure.

[3] Also referred to as 'discounting'

2.1.2 Time pressure

It stands to reason that when participants have less time available to provide causal judgments, behavior will deteriorate in specific ways. In other domains within the larger field of judgement and decision-making the analysis of response time (RT) and the explicit use of time pressure manipulations has led to a better understanding of the cognitive processes involved. Examples include perceptual decision-making (Forstmann et al., 2016; Mulder et al., 2014; Ratcliff et al., 2016), economic decision-making (Couto et al., 2020; Kocher & Sutter, 2006; Reutskaja et al., 2011; Rubinstein, 2007), judgement under uncertainty (Edland & Svenson, 1993; Maule et al., 2000; Ordóñez & Benson, 1997; Young et al., 2012), probabilistic reasoning (Furlan et al., 2016; Gershman & Goodman, 2014), and syllogistic reasoning (Evans et al., 2009; Evans & Curtis-Holmes, 2005).

In contrast, within the causal reasoning literature RT measurements and time pressure manipulations have received little attention. While the effect of time on causal structure learning has been studied a few times (e.g. Coenen et al., 2015; Rehder et al., 2022), we are aware of only one study involving time pressure that directly pertains to causal probabilistic inferences. Experiment 2 in (Rehder, 2014) asked participants to choose under time pressure in which causal network a certain variable value was more likely, but no relevant effects of RT or time pressure were found. Importantly though, it was pointed out that this study can only be considered preliminary (Davis & Rehder, 2020, p. 34); the time pressure manipulation was possibly ineffective and the specific task they used was complex and does not generalize to other paradigms as it required the comparison of two causal configurations.

As time pressure manipulations and RT analysis have enabled significant development in other domains of cognitive science, we here aim to use these methods to spur similar development in our understanding of causal reasoning. Before getting to the application of these methods, let us first discuss some important aspects of them.

2.1.3 Speed-accuracy tradeoff (SAT)

A crucial phenomenon used in the study of time pressure is the speed-accuracy tradeoff (SAT) (Bogacz, Wagenmakers, et al., 2010; Heitz, 2014; Schouten & Bekker, 1967; Wickelgren, 1977). The SAT refers to the common observation that participants can trade accuracy of responding for speed of responding. The typical observation is that faster responses are less accurate. This pattern has been observed across individuals (Bogacz, Hu, et al., 2010; Grice & Spiker, 1979; Miletić & van Maanen, 2019), across conditions (macro-SAT) (Forstmann et al., 2008; Katsimpokis et al., 2020), and within conditions (micro-SAT) (Ridderinkhof, 2002; van Maanen et al., 2018).

While the SAT is often used to refer to the overall accuracy of responses, we can apply the same idea to different ways of measuring accuracy. In the case of causal judgments, we can apply notions of micro- and macro-SAT to the three reasoning errors discussed previously. That is, we can investigate whether the magnitude of these errors changes under an external time pressure manipulation (macro-SAT), and whether they are associated with RTs within conditions (micro-SAT), i.e. the passing of time or ‘internal’ time pressure.

The effects of time pressure, including micro- and macro-SATs, on causal reasoning have not been explicitly studied. Therefore, the main aim of Experiment 1 was to explore how macro- and micro-SAT manifest in a causal judgment task. As Experiment 1 did not address the sources of reasoning errors, we conducted a follow-up experiment, Experiment 2, which included confidence measures to elucidate the underlying processes responsible for reasoning errors.

2.2 EXPERIMENT 1

To test the effects of time pressure on causal probability judgments we used an established causal inference task. In this task participants were asked to make judgments about events that are part of a known causal structure. This 'reasoning from a known structure' entails the applying knowledge of causal relationships to judge the probability of an event conditional on the state of other events in the structure. Besides the implementation of time pressure, the procedure and materials were based on multiple studies by Rehder and colleagues (Davis & Rehder, 2020; Mistry et al., 2018; Rehder, 2018; Rehder & Waldmann, 2017). The experiment was approved by the local ethics committee of the University of Amsterdam (nr. 2019-PML-10019).

2.2.1 Methods

2.2.1.1 Participants

41 individuals participated in the study for course credits (21 female, mean age 21.0) which took 41 minutes on average. All participants provided informed consent before participation in the study. We used an a priori exclusion criterion of an overall mean precision above 18%. As the task uses a percentage response format, this criterion meant that participants with responses on average more than 18 percentage points removed from the normative answer were excluded. This cutoff was chosen as it corresponds to a response strategy in which a participant consistently responds with 50% on all trials and so using it makes sure that the participants included in the analysis did not engage in random responding or guessing. This led to the exclusion of 15 participants. In addition, we removed responses faster than 1.5 seconds, which amounted to 1.8% of responses. In a previous causal reasoning experiment response times ranged between 5.5 and 23 seconds on average (Rehder, 2014). As each trial requires the processing and integration of five cues (three variable values and their causal relationships, see below) responses faster than 1.5 seconds are a clear indication of non-compliance.

2.2.1.2 Experimental design and Procedure

The experiment was conducted in the behavioral sciences lab of the University of Amsterdam. The task consisted of three experimental domains, each consisting of a learning phase and a testing phase. In the learning phase participants learned a specific causal structure, about which they were asked to make inferences in the testing phase. Each testing phase consisted of three blocks with different response deadlines. Each of these blocks consisted of 27 trials. All participants completed 27 trials per domain and deadline condition, for a total of (27 x 3 x 3 =) 243 trials.

The three domains about which participants had to reason concerned meteorology, sociology, and economics (see Rehder, 2014). We tested three 3-variable causal structures (see Figure 2.1): a common cause, a chain, and a common effect structure. Each participant saw all three domains and all three causal structures. The order of the causal structures and which structure was paired with what domain was counterbalanced across participants (e.g. some participants had a common cause structure in the domain of economics, while others had a chain or common effect structure in the economics domain). The variables in the causal structures were binary, each with a "normal" and a non-normal value (Rehder & Waldmann, 2017), which we will refer to in equations with 0 and 1 respectively. The non-normal value for each variable was either "high" or "low" and these values were counterbalanced across participants to control for effects of prior knowledge about the domains (Rehder, 2014).

2.2.1.2.1 Learning phase

Each domain started with a learning phase. First, participants studied several computer screens with verbal information regarding the domain and how the variables are causally related in the specific causal structure. For example, participants that had the economics domain paired with a common effect structure, were taught that low interest rates (cause 1) and high trade deficits (cause 2) independently cause low retirement savings (effect). For the causal relationships it was always the case that non-normal values of variables caused the non-normal value of another variable. The causal structure and relationships was first described to participants in words (e.g.: "High interest rates cause small trade deficits"). For a complete description of all variables and causal relationships used see Appendix A in (Rehder, 2014). After these descriptions participants viewed a graphical representation of the causal network, as in Figure 2.1, but with each node described as the relevant causal variable (e.g. "high interest rates" instead of "X_1", see Figure 2.2).

Next, participants received quantitative information by experience as has been done in previous studies (Rehder & Waldmann, 2017; Rottman & Hastie, 2016). This method involves participants experiencing the causal relationships and their strengths by viewing case data, which is more comparable to how we learn causal information in daily life than to provide participants with written probabilities. For each of the eight possible combinations of variable values participants were presented with a separate screen showing a certain number of cases (each screen corresponded to one row in Table 2.1, see Figure 2.2 for an example of such a screen where all variables have a normal value). The quantitative information regarding the causal networks is learned by the relative number of cases for each possible combination of variable values. This method of teaching participants the network parametrization has been used successfully before (Rehder & Waldmann, 2017; Rottman & Hastie, 2016). The network parametrization of the structures (and thus the number of cases on each sample screen) was taken from Experiment 1a in (Rottman & Hastie, 2016) and intended to be theoretically neutral. The chain and common cause structure had the same parametrization, with base rates[4] of .5 for all variables. The effects in the network had a probability of .75 when their parent was present and .25 when it was not. In the common effect structure, the two causes combined by way of a Noisy-OR gate (Pearl, 1988) with causal strengths of .5 and with base rates of .5. This meant that the effect had 0 probability if no causes were present, .5 when one cause was present, and .75 when both causes were present

[4] We use 'base rate' to refer to the marginal probability of a variable across all cases, e.g. $P(X_1 = 1)$

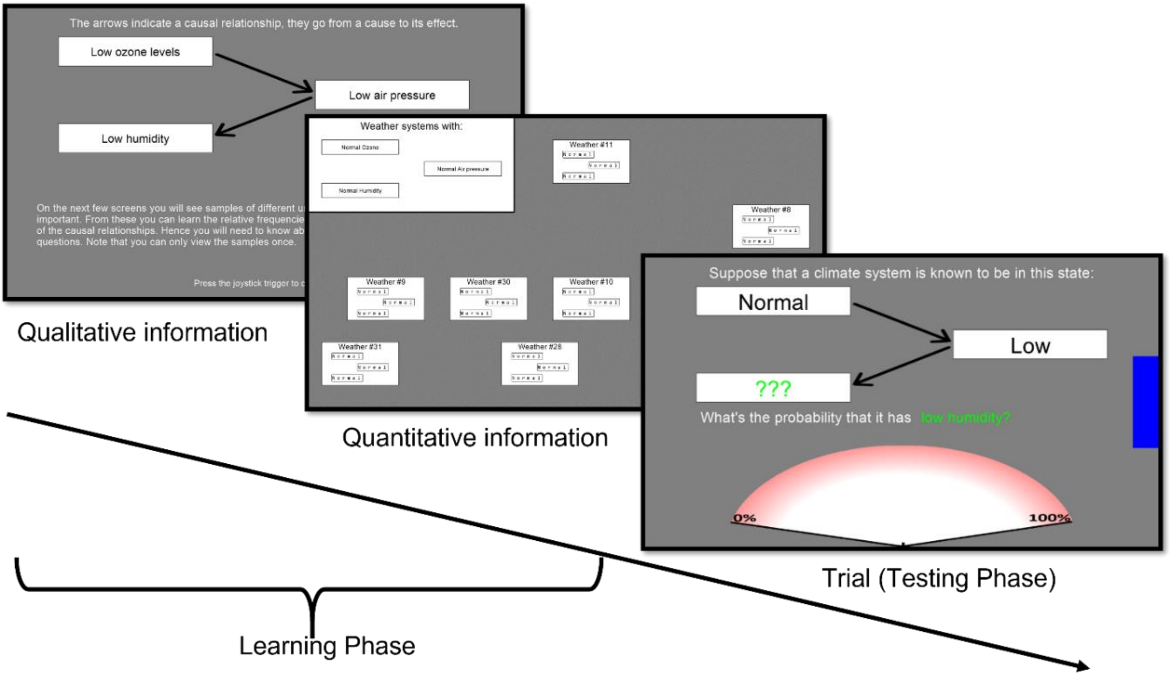
(hence the base rate was .43 for the effect). This parametrization was shown as cases on the sample screens according to Table 2.1.

Causal system state			Number of cases	
X_1	**Y**	**X_2**	Chain and common cause	Common effect
1	1	1	9	6
1	1	0	3	4
1	0	1	1	2
1	0	0	3	4
0	1	1	3	4
0	1	0	1	0
0	0	1	3	4
0	0	0	9	8

Table 2.1 *Parametrization of causal networks used in experiments, implemented as cases viewed by participants. X_1, Y, and X_2 refer to causal variables as presented in Figure 1. In the first three columns the 1s and 0s refer the to presence or absence respectively of a causal variable*

2.2.1.2.2 Testing phase

The testing phase immediately followed each learning phase of a particular domain. Participants judged the probability of a specific variable being in their non-normal state (e.g. "what is the probability of retirement savings being low?"), while the other two variables are presented as in one of three states: unknown, or having their normal or non-normal value (e.g. "trade deficits are normal and interest rates are low", see Figure 2.2). We asked all (3 choices for the queried variable x 3 possible states of first conditional variable x 3 possible states of second conditional variable =) 27 possible questions three times, each under a different level of time pressure, resulting in 81 inferences per domain.



The arrows indicate a causal relationship, they go from a cause to its effect.

On the next few screens you will see samples of different u
important. From these you can learn the relative frequencie
of the causal relationships. Hence you will need to know ab
questions. Note that you can only view the samples once.

Press the joystick trigger to c

Suppose that a climate system is known to be in this state:

What's the probability that it has low humidity?

Figure 2.2 *Overview and screenshots of one domain in Experiment 1. Participants start each domain by learning qualitative and quantitative information about the causal network (panels 1 and 2). The first panel shows a screenshot of how participants learned about the qualitative structure of the causal scenario. The second panel is a screen with learning samples that provided quantitative information. This was one of eight such screens participants saw in each learning phase. Here all variables had the 'normal' value (bottom row Table 2.1), on the 7 other screens the variables had different combinations of values. Each sample was numbered (e.g. 'Weather #23') to emphasize that they represented individual instantiations of the causal variables. The last panel is a trial in the meteorology domain. The blue bar on the right slowly decreased in size indicating the deadline. Participants had to judge the probability of the variable (here 'humidity') with the three green question marks being in its non-normal state (here 'low'). Participants moved a cursor from the center of the part-circle at the bottom of the screen over the edge of the part-circle to indicate their response.*

The testing phase of each domain was split up into three blocks which had response deadlines of 6, 9, and 20 seconds (henceforth referred to as DL6, DL9, and DL20). The choice of deadlines was based on pilot studies, given the lack of experimental findings on the effect of deadlines on causal probabilistic judgments. The presentation order of the blocks was counterbalanced across domains. At the start of each block a screen indicated the response deadline for the next 27 trials. On each trial a blue bar on the right indicated how much time was left to respond (Figure 2.2). When participants failed to answer before the deadline, which happened on 19 trials in total, a screen was presented for 5 seconds before starting the next trial that told participants that they had to respond as accurately as possible while not missing the deadline.

Participants responded on a probability scale ranging from 0% to 100% using a joystick. This setup enabled fast, intuitive responses and reduced variance in RTs due to response execution (see for other uses of joysticks in psychological experiments (Keuken et al., 2015; van Maanen et al., 2012), and for the validity of using a joystick for continuous responses see (Szul et al., 2020)). The 0%-100% range was presented on the edge of part of a circle around the starting point of the cursor controlled by the joystick (see Figure 2.2). Participants were instructed to move the joystick in one swift movement so that the cursor would cross the edge of the circle at the location responding to their answer.

2.2.1.3 Analyses

For our main analyses we applied mixed-effects regression models using the lme4 package in R (Bates et al., 2015). As predictors we included the theoretically relevant variables and their interactions, i.e. RT and deadline condition for all regressions, and for analyses testing specific errors this included variables indicating the state of the causal variables relevant for that error. We included crossed random intercepts for participants and the different inferences (Baayen et al., 2008).

We used linear regressions for RTs and probability judgments, where the judgments were rounded to percentage points. The precision of a judgment was defined as the absolute difference between a response and the normative answer (i.e., lower values indicate a higher precision). Precision is strictly non-negative and positively skewed, and so we used a Gamma distribution with a log link function for the regressions on precision. We added 0.01 to all observations to avoid responses with a precision of 0 (there were 13 such responses in total), as the Gamma distribution is only defined for positive values.

Both probability judgments and precision are coded on a percentage scale (from 0 to 100), and RT is z-scored within participants. As the common cause and chain structures have an identical normative joint distribution (see Table 2.1), both precision and the reasoning errors are measured in exactly the same way (e.g. the inferences relevant to Markov violations are exactly the same). Hence we analyze these structures together.

We report estimates and statistics from regression models with insignificant higher-order interactions removed. To test for this significance of effects we use Satterthwaite adjusted F-tests for the linear mixed regression models (Luke, 2017). Where we applied a Gamma regression for non-normally distributed dependent variables, we use Wald Chi-square tests. Where relevant we also report estimated marginal means, i.e. estimations of the dependent variable based on the regression model for a predictors of interest while averaging over other predictors in the model.

Post-hoc contrasts will be reported with p-values adjusted for multiple comparisons using the Tukey method.

Data and analysis code for the experiments in this paper has been made publicly available at https://osf.io/bz9vj/

2.2.2 Results

2.2.2.1 Manipulation check

To test whether the time pressure manipulation impacted response times we regressed the Deadline factor on RTs and found that the effect of Deadline is indeed significant (*F*(2, 6153) = 668, *p* < .001, Figure 2.3). This is reassuring considering the lack of existing information on time pressure manipulations in this domain.

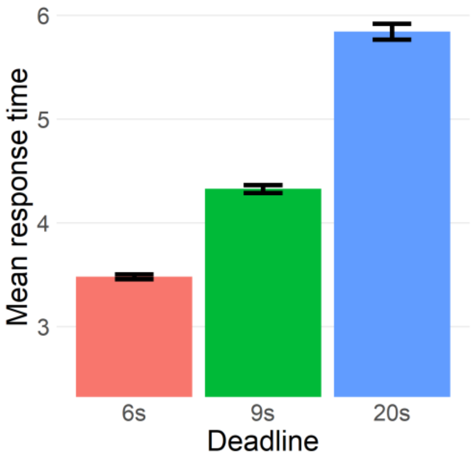


Figure 2.3 *Mean response times per deadline condition. Bars indicate standard error of the mean.*

2.2.2.2 Overall SAT

Next, we investigated the overall SAT, that is, the influence of RTs and time pressure on overall precision. We found a significant main effect of Deadline ($\chi2(2) = 23.8$, $p < .001$), indicating a macro-SAT (Figure 2.4A): Participants were more accurate when there was less time pressure. Post-hoc contrasts revealed that this is due to participants having better precision in the DL20 condition (M = 12.6, SE = 0.680, $z_{DL6\text{-}DL20} = 4.56$, $p < 0.001$, $z_{DL9\text{-}DL20} = 3.54$, $p = .0012$), while we do not find a difference between the DL6 (M = 14.5, SE = 0.818) and DL9 (M = 13.8, SE = 0.737) conditions (z = 1.73, p = .195). We also found a significant interaction effect of RT and Deadline ($\chi2(2) = 8.84$, $p = .012$, Figure 2.4B), revealing a micro-SAT. While the interaction indicated that the effect of RT is less strong with longer deadlines, we found that the effect of RT is significant for each of the deadlines (DL6: M = 1.165, SE = 0.0432, z = 3.72, p < .001; DL9: M = 1.07, SE = 0.0241, z = 3.09, p = .002; DL20: M = 1.03, SE = 0.0131, z = 2.43, p = .015). This means that participants were less accurate the closer they got to the deadline, and that this effect was strongest for the shorter deadlines.

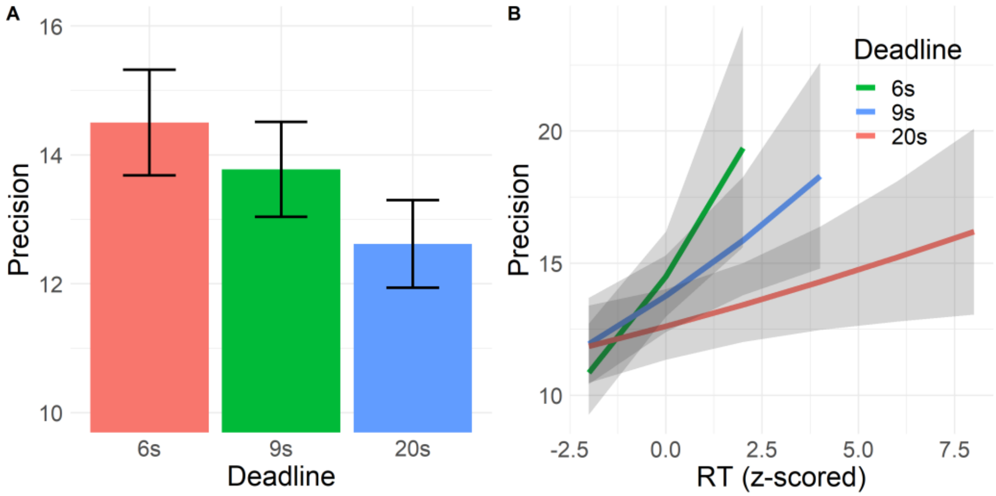


Figure 2.4 *Estimated precision from mixed-effects regression. A. Precision in each deadline condition. Precision is on the Y-axis, defined as the absolute difference between response and normative answer (hence, lower values indicate that participants are more accurate). The bars represent standard errors. B. Interaction effect of deadline and RT on precision. Precision is on the Y-axis, Z-scored RTs are on the X-axis. The lines are estimated marginal means, the ribbons represent the 95% confidence interval.*

2.2.2.3 SAT Markov independence and explaining away

To test the effect of the deadlines and RT on Markov violations we performed another mixed model regression on the data of the common cause and chain networks. Here we analyzed only those inferences relevant to Markov independence, that is, inferences about a terminal variable (X_1 or X_2), while the middle variable was known (either Y=0 or Y=1). The dependent variable was the response in percentage points. We included a variable as fixed effect indicating whether the other terminal event (the screened off variable, X_2 or X_1) was absent, unknown, or present (coded as -1, 0, 1). A significant effect of this factor (henceforth referred to as ScreenedOff) thus indicates a violation of Markov independence. Additional fixed effects for the status of the middle variable (values: Y=0 or Y=1, MidVar), RT, Deadline, and their interactions with ScreenedOff were included in the model.

As expected, we found a significant main effect of ScreenedOff (*F*(2, 1800) = 128, *p* < .001), indicating that participants did not screen off, and thus violated Markov independence. The interactions of ScreenedOff with Deadline (*F*(4, 1801) = 0.641 , *p* = .633) and RT (*F*(2, 1809) =

[5] As we used a log link for these regressions (see Analyses section) on precision, and so these mean effects need to be interpreted multiplicatively (instead of additively in standard regression). Hence a mean effect of RT of 1.16 in the DL6 condition means that if RT increases by 1 SD, the precision is multiplied by a factor of 1.16.

1.68, *p* = .186) were both not significant, indicating that the violations of Markov independence were not impacted by time pressure nor response times (Figure 2.5A and B). While these results (and Figure 2.5) are rather convincing, the lack of a significant effect does not provide direct evidence for the absence of such an effect. As this result is surprising and important for our aims, we additionally computed Bayes factors for the effects of time pressure and RT on reasoning violations and conservatism, both here and in subsequent sections. Bayes factors provide strong evidence against an effect of both deadlines (BF_{01} > 100) and RT (BF_{01} = 23.0). We did find a significant interaction between ScreenedOff and MidVar (*F*(2, 1801) = 15.7, *p* < .001)), indicating that the violations of Markov dependence were larger when the middle variable was present than it was not (Figure 2.5A and B).

We performed a similar analysis for the common effect structure, the only difference being that the model did not include the variable MidVar. Hence the initial model included fixed effects for ScreenedOff, Deadline, RT, and the interactions of the latter two with ScreenedOff. We again only found a significant main effect of ScreenedOff (*F*(2, 415) = 29.4, *p* < .001), indicating that participants violated Markov independence here. The interactions with ScreenedOff were not significant for both Deadline (*F*(4, 415) = 1.65, *p* = .161, BF_{01} = 5.85) and RT factors (*F*(2, 419) = 0.0105, *p* = .9896, BF_{01} = 24.0, Figure 2.5C).

To test the effect of the deadlines and RT on explaining away we conducted another mixed model regression using only the data from inferences relevant to explaining away. That is, those inferences in the common effect structure about one of the causes, while knowing that the effect (the middle variable) is present. We computed a variable (AwayVar) indexing whether the other terminal event (X_j above) was absent, unknown, or present (coded as -1, 0, 1 respectively). Next, we recoded participants' responses on these trials such that in our model the normative explaining away pattern would result in the effect of AwayVar being zero[6]. This means that we can interpret the effect of AwayVar as deviations from the normative pattern of explaining away. Our model included AwayVar, Deadline, RT, and the interactions of Deadline and RT with AwayVar.

We found a significant main effect of AwayVar (*F*(2, 426) = 432, *p* < .001), indicating that participants did not engage in the normative explaining away pattern (clearly visible in Figure 2.5D). The effect of knowing that the other cause was absent compared to it being unknown is +0.46% (*SE* = 2.80), which is far from the CBN prediction, which says that the probability should increase by 28.6% compared to when the state of the other cause is unknown. The effect of knowing that it is present is +4.88% (*SE* = 2.89), which again is far from the CBN prediction of -11.4%.

[6] The CBN predictions for P(Xi|Y=1, Xj=0), P(Xi|Y=1), and P(Xi|Y=1, Xj=1) are 1, .714, and .6 respectively. We recoded participant's responses by substracting .286 from the inference where Xj=0, and adding .114 to the responses on Xj=1. Hence in this recoded format the normative responses for all three inferences are .714 and the effect of AwayVar is zero. In the case of a non-normative explaining away pattern, we would find a nonzero effect of AwayVar.

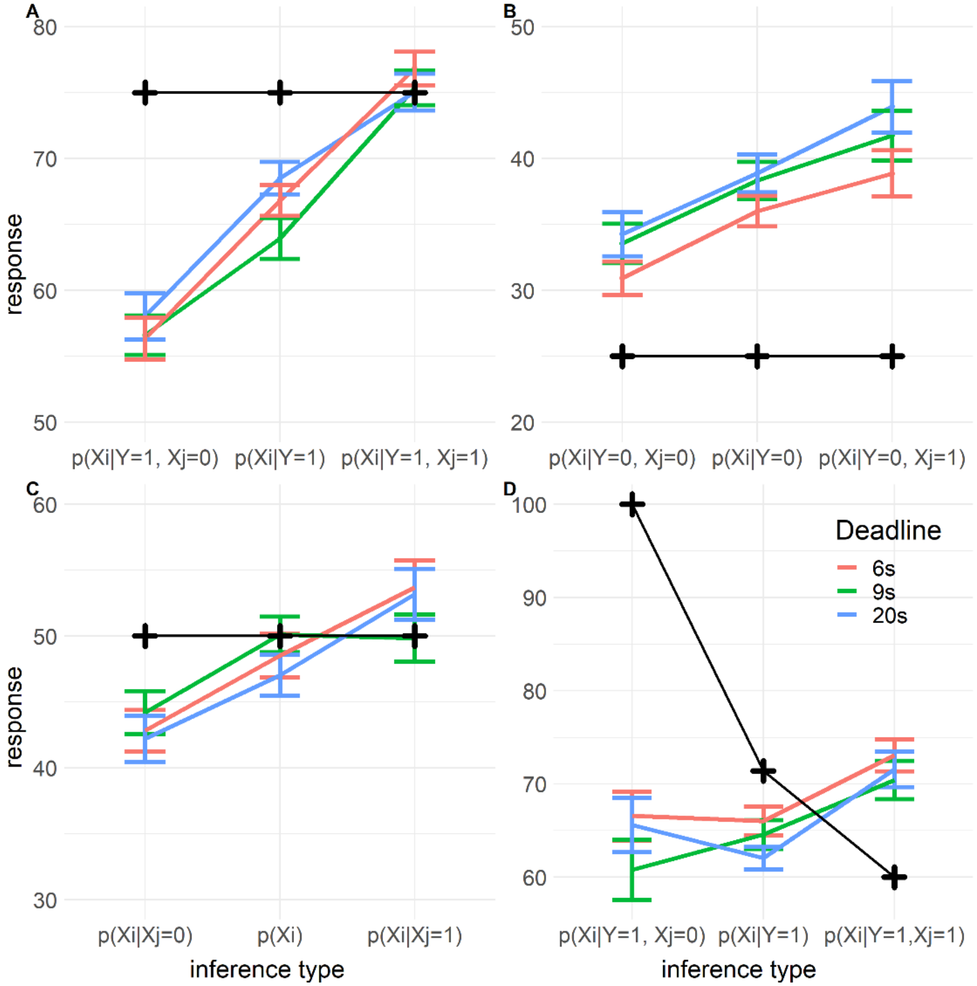


***Figure 2.5** Markov violations and explaining away per deadline in Experiment 1. Y-axis indicates response on a percentage scale. Colored lines indicate mean responses, the error bars indicate their standard errors. The black crosses indicate the normative response. The x-axis indicates the specific inference. Symmetric inferences are collapsed, e.g. ‘P(Xi | Xj=0)’ refers to both ‘P(X_1=1| X_2=0)’ and ‘P(X_2=1| X_1=0)’. A. Markov violations in common cause and chain structures where the middle variable is present (Y=1). B. Markov violations in common cause and chain structures where the middle variable is absent (Y=0). C. Markov violations in common effect structure. D. Explaining away in common effect structure.*

There was no influence of deadlines on how participants explained away ($F(4, 426) = 1.27$, $p = .280$, $BF_{01} = 11.0$). However, we did find mixed evidence of an interaction of AwayVar with RT ($F(2, 433) = 3.64$, $p = .0270$, $BF_{10} = 1.09$). We plotted the estimated interaction in Figure 2.6. From this interaction we can see that RT impacted responses on trials where the non-queried cause is present ($t(435) = -3.36$, $p < .001$), while RT had no effect when the non-queried cause is absent ($t(432) = -0.79$, $p = .43$), or when its status is unknown ($t(434) = -0.21$, $p = .83$). The responses on trials where the non-queried cause is present got closer to 50% percent as participants took longer to respond. It is possible that this effect is not related to explaining away, but rather to conservative inferences. The inference where the RT has an effect is where responses are most extreme and so we would expect conservatism to be more pronounced for this inference. We return to this in the discussion.

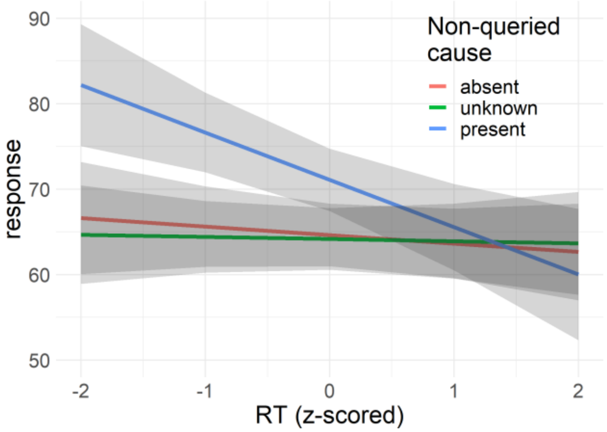


Figure 2.6 *Interaction of Explaining Away with response times. Estimated responses from mixed-effects regression in common effect structure based on RT and AwayVar in Experiment 1. This plot visualizes the interaction effect AwayVar x RT on responses, i.e. the interaction between RT and explaining away, where horizontal lines would indicate no effect of RT and differently sloped lines indicate an interaction. AwayVar here refers to the status of the non-queried cause in the inference on the common effect structure, where -1 indicates the non-queried cause is absent, 0 that its status is unknown, and 1 that the non-queried cause is present. The y-axis indicates responses on percentage scale, the x-axis indicates RT (z-scored). The grey ribbons represent 95% confidence intervals.*

2.2.2.4 SAT conservative inferences

Participants tended to respond conservatively with responses being on average 4.9 percentage points (SE = 0.84, t = 5.85, p < 0.001) closer to 50% than the normative response (see Figure 2.7). To quantify the relationship between time pressure and conservative responses, we computed a variable that measured the distance that a response moved from the normative answer towards 50% for all conflict (e.g. $P(Y = 1|X_i = 1, X_j = 0)$) and ambiguous (e.g. $P(Y = 1|X_i = 1)$) inferences. Positive values for this variable indicate that a response was in between 50% and the normative answer (or at 50%). Positive values thus indicate conservative inferences. Negative values indicate that a response was more extreme (closer to 0% or 100%) than the normative response. Because this variable cannot represent responses for which the normative response is exactly 50% these trials were excluded from this analysis (21.9% of all trials). Additionally, we removed responses on which participants indicated a probability that was in the oppositive half of the measurement scale as the normative response (12.6% of all trials). For example, if the normative response was larger than 50%, but a participant gave a response below 50%, that trial was removed.

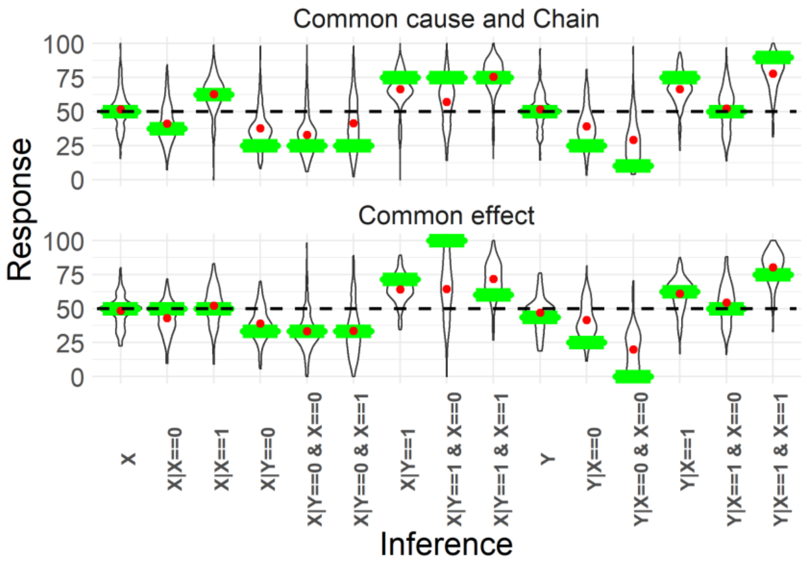


Figure 2.7 *Mean responses per inference. This figure indicates conservatism in both Common cause and Chain (top) and Common effect structures (bottom). Y-axis represents the response (in %), X-axis indicates the type of inference. We collapsed over the symmetry between X_1 and X_2. The violin plots indicate the response distribution of all participants, the red dot is the mean response. The green bars indicate the normative (CBN) response. The horizontal dashed black line indicates responses at 50%. Conservatism can be seen by red dots that are closer to 50% than the green bars (for inferences where the normative response is not 50%).*

To test the impact of time pressure on conservatism we employed a regression model using the metric of conservative responding as dependent variable. We found a significant interaction effect of Deadline and RT on conservatism (F(2, 2673) = 6.89, p = .001, BF10 = 6.48, Figure 2.8). Using post-hoc contrasts, we found that the effect of RT is significant in the 6s (β = 1.90, SE = 0.679, t(2676) = 2.79, p = .0054) and 9s deadlines (β = 1.49, SE = 0.378, t(2672) = 3.95, p < 0.001), but not for the 20s deadline (β = 0.132, SE = 0.215, t(2672) = 0.612, p = .54). Pairwise contrasts revealed that the effects in the 6s and 9s conditions are not significantly different (t(2672) = 0.520, p = . 86), while they were different from the 20s condition (versus 6s: t(2676) = 2.48, p = .036; versus 9s: t(2672) = 3.14, p = .0049). Hence there seemed to be a micro-SAT for conservative inferences in the 6s and 9s conditions, but not in the 20s condition. This is in line with an overall main effect of RT (F(2, 2678) = 18.5, p < .001, BF10 = 4.59). We found mixed evidence for a main effect of Deadline (F(2, 2668) = 4.40, p = .012, BF01 = 9.34) which could indicate a macro-SAT. Contrasts indicate that there is more conservatism in the 6s deadline condition (M = 6.39, SE = 1.43) compared to the deadlines of 9s (M = 4.93, SE = 1.39, t(2670) = 2.74, p = .017) and 20s (M = 4.95, SE = 1.39, t(2670) = 2.67, p = .020).

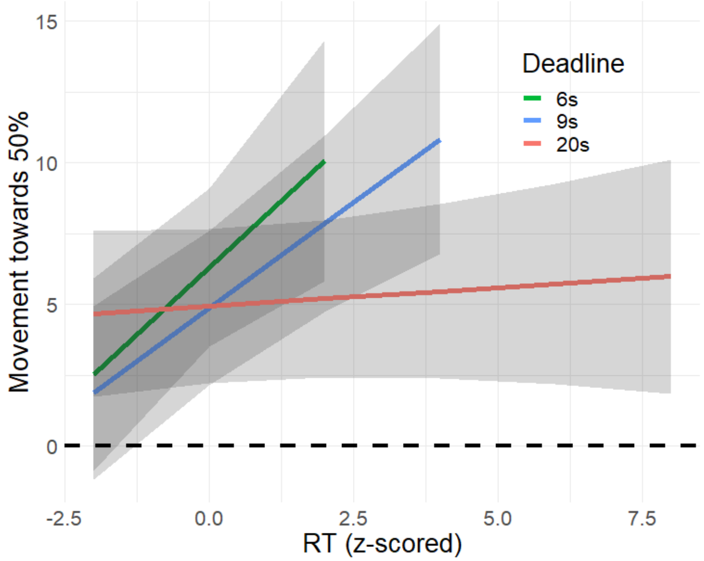


Figure 2.8 *Estimated movement towards 50% based mixed-effects regression in Experiment 1. This plot visualizes the Deadline x RT interaction on conservatism. The y-axis represents conservatism, that is, the distance a response moved from the normative answer towards 50%. The x-axis represents RT (z-scored). The colored lines indicate the predictions, separated per deadline. The grey ribbons represent 95% confidence intervals.*

2.2.3 Discussion

Taken together, the results from Experiment 1 indicate that there is an overall macro-SAT in causal probability judgements. Time pressure decreases the accuracy of responses as compared to the normative CBN point prediction. In addition, we found evidence for a micro-SAT. Responses with longer RTs are generally less accurate, and this micro-SAT is stronger in the conditions with more time pressure.

Moreover, the results of Experiment 1 reveal that Markov independence violations are not impacted by time pressure. Of the inferences relevant for explaining away, the only effect of time pressure we found was an effect of RT on the inference where the non-queried cause is present (Figure 2.6). However, it seems that this is due to conservative responding, as we do not see the effect on the other inferences relevant for explaining away. For the single inference where we find an effect of RT participants provided estimates closer to 50% when RTs were longer. This could indicate that participants grasped the idea that the non-queried cause being present should not increase the probability of the queried cause when they took more time to respond, i.e. participants grasped explaining away when they took more time. If this were the case, however, the question remains why we do not see the opposite effect of RT for those trials where the non-queried cause is absent, which we would expect if the effect of RT is related to grasping the idea of explaining away. Therefore, a more plausible explanation for the effect of RT is not related to explaining away, but due to the phenomenon of conservative inferences. If the passing of time affects conservative responding, we would most clearly see this on trials where responses are farther away from 50% and this is what we find here.

This conjecture is bolstered by the fact that we do find that conservative responding is impacted by time: Both the deadline manipulation as well as the passing of time increase conservative responding. This latter effect is strongest in the conditions with short deadlines. This is the same pattern found for overall precision, and so it seems that changes in accuracy associated with time pressure and RTs are due to changes in conservative responding.

The observation that conservative responding is differently related to time pressure than Markov violations and failures to explain away is an indication of different cognitive processes. In particular, an interesting hypothesis is that the conservatism is the result of an increased probability to respond at or near 50% when time increases, effectively guessing (van Maanen, 2016). This would entail that slow responses are associated with low confidence, reflecting that participants were unsure of those answers (Rahnev et al., 2020). Hence conservative responding might not be an error specific to causal reasoning, but the result of a more general cognitive principle related to uncertainty. This hypothesis is tested in Experiment 2.

2.3 EXPERIMENT 2

Experiment 2 tests the hypothesis that the increase in conservative inferences under time pressure is related to a decrease in confidence. In addition, Experiment 2 serves as a replication of Experiment 1, which seems opportune given the scarcity of experimental findings on time pressure effects in causal judgments. Moreover, there was a sizeable dropout rate (37% of participants) in Experiment 1 due to the a priori threshold we set on the precision of responses. Such a sizable dropout is not uncommon for demanding causal reasoning tasks like ours. For example, in a set of experiments on causal attribution the dropout rate ranged from 29% to 44% (S. G. B. B. Johnson & Keil, 2014), in a set of studies on diagnostic inference it has been consistently around 30% (Meder et al., 2014; Meder & Mayrhofer, 2017), and a set of experiments on the effect of prescriptive norms on causal inferences had dropout rates of up to 37% (Samland & Waldmann, 2016). But, while a dropout rate like in Experiment 1 is not uncommon, it still behooves us to replicate the findings.

2.3.1 Confidence

Besides replication another goal for Experiment 2 was to study the relationship between confidence and conservatism. Confidence has been considered an important component of reasoning and decision-making more generally (Ackerman & Thompson, 2017; Pleskac & Busemeyer, 2010; Rahnev et al., 2020; Ratcliff & Starns, 2013; N. Yeung & Summerfield, 2012). Confidence tends to correlate negatively with RT, and higher confidence is associated with more accurate judgments (Rahnev et al., 2020). An important consequence of employing time pressure manipulations is that participants need to make responses with varying levels of confidence (Hoge, 1970; Pleskac & Busemeyer, 2010; Vickers & Packer, 1982). Which we also expect to observe. With regard to causal reasoning, one recent study has shown that confidence can predict verbal causal ratings (O’Neill et al., 2022), indicating that we can expect confidence to be relevant in understanding causal judgments. Additionally, confidence has been used to test theories of memory retrieval (Ratcliff et al., 1995; Ratcliff & Starns, 2013), sensory processing (Green &

Swets, 1966), and decision-making (Balakrishnan, 1996; Mueller & Weidemann, 2008), suggesting that confidence ratings can indeed reflect differences in cognitive processing.

2.3.2 Methods

2.3.2.1 Participants

37 individuals participated in the study (9 female, mean age 28.3) for a monetary reward of £5.63 via the Prolific platform (www.prolific.co). We selected participants that had a 100% approval rating for previous studies they participated in on Prolific. All participants provided informed consent and the experiment took around one hour to complete. We used the same exclusion criteria as in Experiment 1. This resulted in the exclusion of 20 participants due to a mean precision above 18% (we return to this in the discussion), and the removal of 2.2% of responses with an RT of lower than 1.5 seconds.

2.3.2.2 Design and Procedure

The experimental design of Experiment 2 was almost identical to Experiment 1, but differed from it in three ways: (1) the experiment was conducted online rather than in a physical laboratory, (2) participants responded using their mouse or trackpad instead of a joystick, and (3) at the end of each trial participants reported their confidence. The use of a mouse or trackpad required an additional screen after each response where participants were presented with their current cursor position and were asked to move it back to the middle of the screen. The crucial difference between the experiments was that, in addition to participants' probability estimates, we now also asked participants after each trial to report the confidence they had in their response. Identical to the probability judgments, participants moved a cursor over the edge of part of a circle, which here ranged from 'not confident at all' to 'completely confident'.

2.3.2.3 Analyses

We used the same mixed effects regression approach as in Experiment 1. For the analyses regarding confidence we added z-scored confidence reports as an additional predictor.

2.3.3 Results

2.3.3.1 Replication of SAT findings from Experiment 1

We largely replicated the effects on response precision and reasoning errors from Experiment 1. Hence, we only briefly report here the main results related to the systematic deviations from CBN predictions (additional analyses are reported in Appendix A). We found no effect of deadlines ($F(4, 1171) = 0.960$, $p = .43$, $BF_{01} = 48$, Figure 2.9A and 9B) or RT ($F(2, 1180) = 2.73$, $p = .065$, $BF_{01} = 3.13$) on Markov independence violations in common cause and chain structures, nor did we find an effect on these violations in the common effect structure (Deadline: $F(4, 262) = 1.08$, $p = .37$, $BF_{01} = 9.93$, Figure 2.9C; RT: $F(2, 265) = 1.73$, $p = .18$, $BF_{01} = 3.89$). For explaining away we found no effect of deadlines ($F(4, 271) = 1.18$, $p = .32$, $BF_{01} = 11.7$, Figure 2.9D), but similar to Experiment 1 we found an indication of an effect of RT ($F(2, 279) = 6.29$, $p = .002$,

BF_{10} = 0.986). This latter effect seems again to be due to conservative responding as in Experiment 1. We found that the effect of RT is significant for the inferences where the non-queried cause is present ($P(X_i = 1|Y = 1, X_j = 1)$, β = -6.58, *SE* = 1.44, *t*(280) = -4.56, *p* < .001) or unknown ($P(X_i = 1|Y = 1)$, β = -4.30, *SE* = 1.64, *t*(275) = -2.63, *p* = .009). For these inferences participants responded closer to 50% when RTs were longer. There was no effect of RT on the inferences where the non-queried cause is absent ($P(X_i = 1|Y = 1, X_j = 0)$, β = 1.77, *SE* = 1.57, *t*(276) = 1.12, *p* = .263). Lastly, we found mixed evidence of an interaction of Deadline and RTs on conservative responding (*F*(2,1676) = 4.55, *p* = .011, BF_{10} = 0.739). Focusing on main effects, we find that there is no effect of Deadline on conservatism (*F*(2,1673) = 1.93, *p* = .15, BF_{01} = 21.3), but we find a large effect of RT (*F*(1,1681) = 21.5, *p* < .001, BF_{10} > 100) indicating that conservatism is sensitive to internal time pressure.

Having replicated the main findings of Experiment 1, let us now look at confidence as a predictor of precision and reasoning errors.

2.3.3.2 Role of confidence in overall precision

Regression analysis showed there was a main effect of confidence on precision (χ2(1) = 24.0, p < 0.001, β = -6.14, SE = 1.25), indicating that more precise responses were associated with higher confidence. This result gives credence to the use of confidence as an index of participant's uncertainty about their inference. We also found an interaction effect of confidence and RT on precision (χ2(1) = 5.15, p = .023, β = -2.62, SE = 1.16). For responses associated with low confidence longer RTs imply worse precision, while for responses with high confidence longer RTs imply better precision. The most likely interpretation for this finding is that we find longer RTs for two reasons. Sometimes long RTs reflect more deliberation, leading to more accurate responses and higher confidence. But sometimes long RTs reflect that the problem is difficult, resulting in less accurate and less confident responses. There was no interaction of confidence with the deadlines (χ2(2) = 0.822, p = .66). In addition, the main effect of Deadline on precision remains significant (χ2(2) = 6.47, p = .0394).

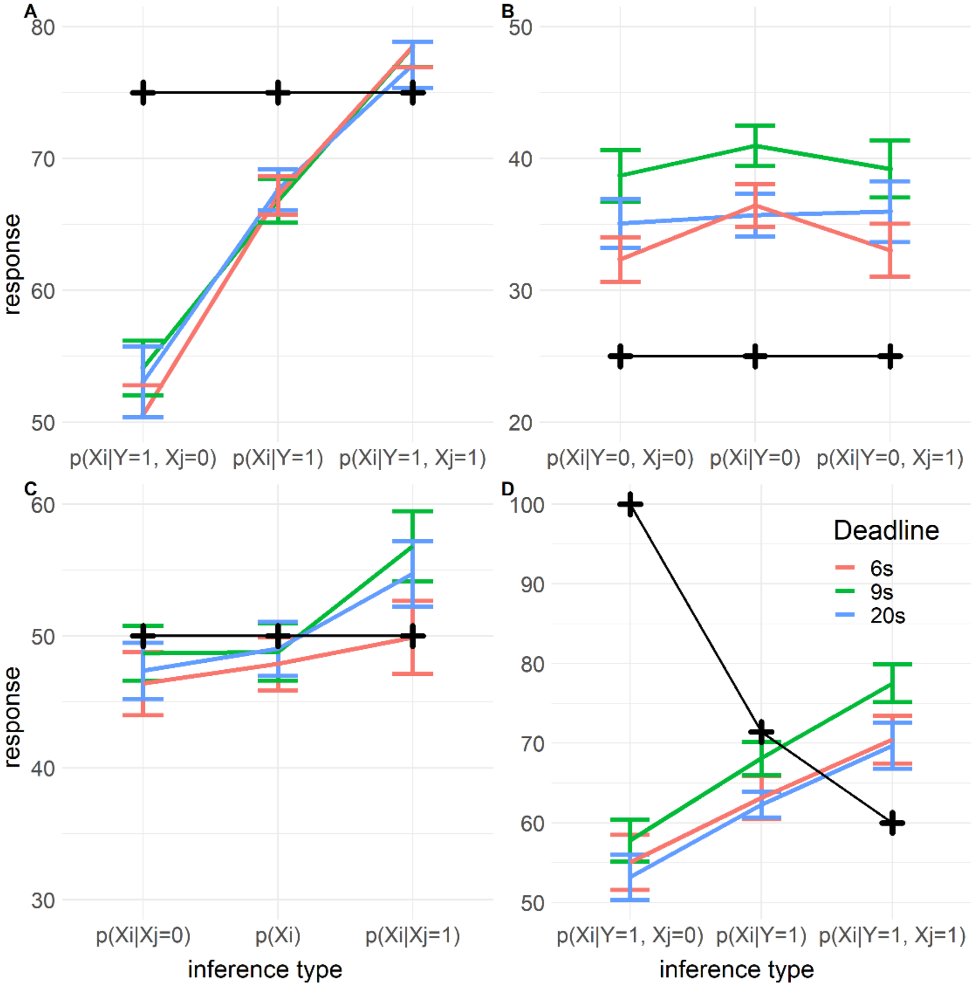


Figure 2.9 *Markov violations and explaining away per deadline in Experiment 2. Y-axis indicates response on a percentage scale. Colored lines indicate mean responses, the error bars indicate their standard errors. The black crosses indicate the normative response. The x-axis indicates the specific inference. Symmetric inferences are collapsed, e.g. 'P(Xi | Xj=0)' refers to both 'P(X1 | X2=0)' and 'P(X2 | X1=0)'. A. Markov violations in common cause and chain structures where the middle variable is present (Y=1). B. Markov violations in common cause and chain structures where the middle variable is absent (Y=0). C. Markov violations in common effect structure. D. Explaining away in common effect structure.*

2.3.3.3 Role of confidence in Markov violations and explaining away

To test the role of confidence in Markov violations and explaining away we included confidence as an additional predictor in our regression analyses and inspected its interactions with the violations. For the common cause and chain structures we find a significant three-way interaction effect of confidence with the screened off variable (ScreenedOff) and the status of the middle variable (MidVar; *F*(2, 1160) = 6.21, *p* = .003, BF_{10} = 1.99) as well as a significant two-way interaction of confidence with ScreenedOff (*F*(2, 1164) = 4.13, *p* = .016, BF_{10} = 4.78). To understand the relationship between these Markov violations and confidence we plotted the response estimates at different levels of confidence in Figure 2.10. Firstly, post-hoc testing indicates that there is no violation of Markov independence when the middle variable is absent (Y=0), as there are no differences between the levels of ScreenedOff ($\Delta_{\text{ScreenedOff -1 vs 0}}$ = 1.86, *SE* = 1.48, *t*(1155) = 1.26, *p* = .419; $\Delta_{\text{ScreenedOff 0 vs 1}}$ = 3.04, *SE* = 1.49, *t*(1155) = 2.04, *p* = .103). In the case when the middle variable was present (Y=1), there were significant Markov violations ($\Delta_{\text{ScreenedOff -1 vs 0}}$ = 12.6, *SE* = 1.43, *t*(1155) = 8.23, *p* < .001; $\Delta_{\text{ScreenedOff 0 vs 1}}$ = -5.36, *SE* = 1.74, *t*(1159) = -3.09, *p* = .0059). For this case where the middle variable was present, we can see from the slopes of the colored lines in Figure 2.10A that the Markov violations were stronger when confidence is high (the red lines are the steepest) when comparing the inferences where X_j is absent versus when it is unknown. There seems no change in magnitude of the violation comparing the inference where X_j is unknown versus when it is present (colored lines have the same slope). This is confirmed by looking at the effect of confidence on these inferences, as the effect is larger when Xj is unknown compared to when it is absent (Δ = 4.04, *SE* = 1.64, *t*(1163) = 2.47, *p* = .037, first two columns Figure 2.10A), but it is not different when compared to when X_j is present (Δ = 0.107, *SE* = 1.65, *t*(1163) = 0.065, *p* = .998, second and third columns Figure 2.10A). So higher confidence seems to lead to a larger independence violation exclusively when comparing the $P\left(X_i = 1 \middle| Y = 1, X_j = 0\right)$ and $P(X_i = 1|Y = 1)$ inferences, and seemingly not for the other inferences (Figure 2.10).

We did not find an effect of confidence on Markov violations in common effect structure (*F*(2, 263) = 0.123, *p* = .88, BF_{01} = 11.0), nor did we find an effect on explaining away (*F*(2,273) = 0.563, *p* = .57, BF_{01} = 15.4). All in all, it seems that Markov violations or failures to explain away are not systematically related to the confidence participants have in their responses.

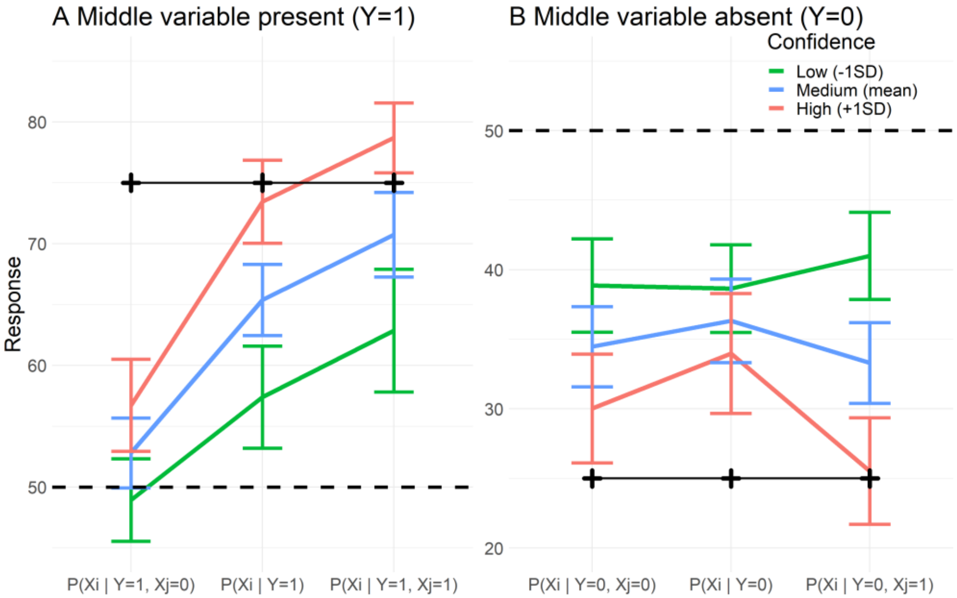


Figure 2.10 *Effect of confidence on Markov violations in common cause and chain structures. Colored lines are estimated responses based on regression model, error bars indicate 95% confidence interval. The estimates are based on different levels of confidence, low (-1SD), medium (mean), and high (+1SD). Black crosses and solid black lines indicate normative answers. Dashed line indicates 50%. This plot visualizes the three-way interaction effect of confidence with the screened off variable (ScreenedOff) and the status of the middle variable (MidVar).*

2.3.3.4 Role of confidence in conservative inferences

To analyze the role of confidence in conservative inferences we again used the same variable as in Experiment 1 for how much a response moved from the normative answer towards 50%. Figure 2.11A plots participants responses against confidence, which indicates that more extreme responses were associated with higher confidence.

We conducted a regression on how far responses moved towards 50% using RT, Deadline, confidence and all interactions as predictors. We found mixed evidence for a three-way interaction ($F(2, 1668) = 3.16$, $p = .043$, $BF_{10} = 0.328$, Figure 2.11B). From Figure 2.11B it seems participants responded more conservatively the closer they got to a deadline, and this effect is strongest for the shorter deadline conditions. Tentatively, this effect seems to be moderated by confidence. Responses with low confidence were generally already conservative regardless of RT and time pressure, while those responses with high confidence moved further towards 50% the closer they got to the deadline (see Figure 2.11B). Most importantly, we found strong evidence for a main effect of confidence ($F(1, 1675) = 236$, $p < 0.001$, $BF_{10} > 100$, $\beta = -3.46$, $SE = 0.436$), indicating that responses with lower confidence tended to be closer to 50%. This last point is visible from the colors in Figure 2.11A, which indicate how much responses moved towards 50%.

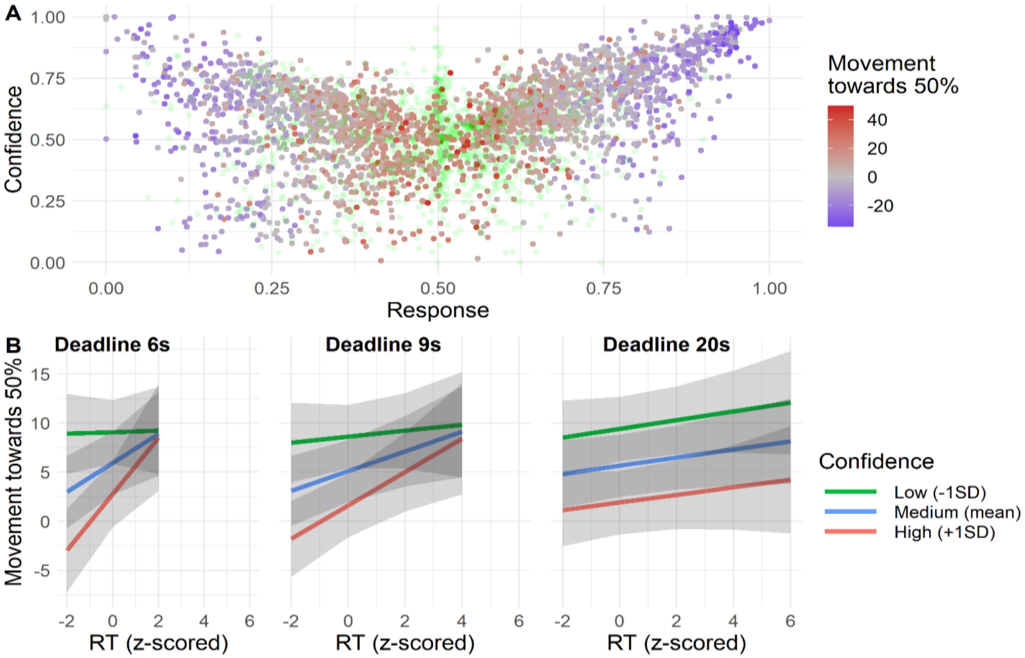


Figure 2.11 *Confidence and conservatism. A. Scatterplot of responses and confidence. Responses are colored from red through grey to blue based on the distance they moved towards 50% from the normative answers (i.e. conservatism). Green transparent dots represent responses that either moved beyond 50% or for which the normative answer was 50%, and so cannot be color-coded based on their movement towards 50%. B. Plots of interaction effect of deadline, RT, and confidence on response movement towards 50%. Based on regression model discussed in text. Grey ribbons indicate 95% confidence interval.*

2.3.4 Discussion

We replicated the main findings of Experiment 1 in an online replication experiment. We again found that Markov violations and failures to explain away are not impacted by time pressure. We did find an effect of RTs on explaining away, but this, like in Experiment 1, seems to be better explained by conservative responding. Participants decreased their response, getting closer to 50%, with longer RTs on the inferences where the non-queried cause was present and when it was unknown. Conservatism on these inferences should be more pronounced than for the inference where the non-queried cause is absent as the mean responses are more extreme for the former. Additionally, as the mean responses to the inference were below the normative response (Figure 2.9C), the effect of RT means that these responses are further away from the normative response with longer RTs. If the effect of RT was due to participants more properly explaining away, we would expect this inference to increase, which is the opposite of what we find.

We also replicated the effects of time pressure on conservatism. Participants respond more conservatively when pressure to respond increases (i.e. they approach the temporal deadline). In agreement with Experiment 1, it seems that conservative responses are qualitatively different from Markov violations and failures to explain away. This conclusion is corroborated by our analysis of confidence reports, which showed that low confidence is associated with conservative responding, while not being systematically related to Markov violations or failures to explain away. We did find an effect of confidence on Markov violations but only when comparing the $P(X_i = 1|Y = 1, X_j = 0)$ and $P(X_i = 1|Y = 1,)$ inferences in the common cause and chain structures (Figure 2.10). The difference between these inferences was larger when confidence was high. However, this effect was not present for the other inferences in the common cause and chain structures, nor for the Markov related inferences in the common effect structure. If higher confidence was truly related to larger Markov violations we would expect the effect of confidence to be present for all Markov violations, not just the aforementioned two inferences. The only systematic relationship between confidence and the inferences related to Markov violations seems to be that higher confidence responses are more extreme (see Figure 2.10).

We established that more conservative judgments were systematically related to low confidence, regardless of time pressure (see Figure 2.11B). Responses high in confidence, however, seemed to vary in conservatism (red lines in Figure 2.11B). We return to this in the general discussion.

It is notable that there was a substantial drop-out rate as in Experiment 1 due to the precision exclusion criterion. As discussed in the introduction to Experiment 2, while such dropout rates are not uncommon for demanding reasoning tasks, it was a reason for us to replicate the findings from Experiment 1. Experiment 2 had a larger dropout rate than Experiment 1, which can be expected given the possibility of lower task compliance when conducting online experiments (Chandler et al., 2014; Crump et al., 2013; Dandurand et al., 2008; Paré & Cree, 2009). Nevertheless, we replicated the main findings from Experiment 1 using an online task. Moreover, as the participants included in the analysis have a precision below 18% due to the a priori exclusion criterion, we know that they are performing above chance and so understand the task to a certain degree. Thus, at minimum, our findings are stable for the subpopulation of people that certainly understand the task and comply with task instructions.

2.4 GENERAL DISCUSSION

Our experiments were aimed at elucidating the cognitive effect of time pressure on causal reasoning. To this extent we asked participants to draw causal inferences from known causal structures and manipulated the available time to respond while measuring RT.

We found that time pressure led to quicker and less accurate responses, i.e. we found an overall macro-SAT, in line with numerous studies on other types of reasoning and decision-making (Bogacz, Wagenmakers, et al., 2010; Heitz, 2014). We also established that overall performance decreased with response time, a micro-SAT (Heitz, 2014). Together this means that while participants were overall less accurate when presented with shorter deadlines, within each deadline condition their least accurate responses were those that took the longest amount of time.

Additionally, we investigated the effect of time pressure on persistent patterns of non-normative responding: Markov violations, failures to explain away, and conservative inferences (Rehder, 2018; Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016). We found that overall the magnitudes of neither Markov violations nor failures to explain away are affected by time pressure, neither on the macro- nor on the micro-SAT level. This is in line with the conclusion in (Rehder, 2014) that such violations can be the result of careful and deliberative reasoning.

In contrast, the magnitude of conservatism is impacted by time pressure. These conservative responses were more common under stronger time pressure and conservatism increased when participants approach the response deadline. This effect is stronger for shorter deadlines. As this response pattern appears similar to the overall SAT effects, it is plausible that the overall SAT is due to changes in the amount of conservatism. This is corroborated by our analysis of confidence in Experiment 2, where we find that both conservative responding and overall accuracy are related to participants' confidence in their judgments. Conservatism in responding was most severe when participants were uncertain about their judgments. Hence, we conjecture that the pattern of conservative responding we found is due to uncertainty and experienced time pressure. This would explain why participants respond more conservatively the closer they get to a deadline, and why this effect is substantially smaller in the longest deadline condition, where participants rarely responded close to the deadline.

2.4.1 Potential explanations of causal reasoning

What are the implications of these findings for existing theoretical accounts? The first thing to note is that the finding that Markov violations and failures to explain away are not sensitive to time pressure is surprising for a number of reasons. Firstly, accuracy in judgement and decision-making typically decreases under time pressure (Bogacz, Wagenmakers, et al., 2010; Heitz, 2014). Secondly and more importantly, it seems that promising explanations of exactly these violations predict that they would increase; this includes the Mutation Sampler (Davis & Rehder, 2020), and heuristic-based explanations (Rehder, 2014; Rehder & Waldmann, 2017; Rottman & Hastie, 2016) including the Quantum Probability theory (Trueblood et al., 2017).

2.4.1.1 Sampling theory

Davis and Rehder (Davis & Rehder, 2017, 2020) proposed a theory of causal reasoning that accounts for the normative violations based on a sampling procedure. This work built upon recent developments in cognitive science that proposes sampling schemes to underlie a variety of probabilistic judgements in different domains (Dasgupta et al., 2017; Hertwig & Pleskac, 2010; Lieder et al., 2012; Vul et al., 2014). The model, termed Mutation Sampler, proposes that people engage in sampling over states of a causal network to make inferences. Subsequently people compute inferences based on the relative frequencies of events in the generated samples.

The Mutation Sampler mechanism posits that Markov violations and failures to explain away are due to a biased and limited sampling procedure. Deadline-induced time pressure would further limit this sampling procedure, making the bias more pronounced, leading to larger violations. Similarly, longer RTs would indicate a longer sampling procedure, reducing bias and hence these violations. Neither of these predictions is borne out in our experiments.

Consistent with our findings, the Mutation Sampler does predict an increase in conservative responding due to time pressure. The prediction is not directly due to the bias in sampling, but relates to the probability of sampling from the causal states necessary to calculate the required relative frequency. If these states are not sampled, the Mutation Sampler predicts a default response of 50%. When presented with less sampling time (e.g., due to a response deadline), the number of trials on which these critical states are not yet sampled increases. This would result in an increase of responses centered at 50%. In both our experiments we find spikes at 50%;

participants responded with 50% on 5.1% of all trials. To test if the spikes change due to the deadlines we conducted a repeated-measures ANOVA using the frequency of response between 49.5% and 50.5% as dependent variable. We find that the deadlines have no effect on the size of the spikes ($F(2,84) = 1.46$, $p = .239$, $BF_{01} = 4.03$). This suggests that the effect of time pressure on conservative responding cannot be attributed to an increase of responses at 50%. It should be noted however that the Mutation Sampler is not the only possible implementation of a sampling approach to causal reasoning (Davis & Rehder, 2020). For instance, the Mutation Sampler could be modified by incorporating a prior probability distribution that weights responses near 50% more strongly, which would be able to predict conservatism not just by spikes of responses at 50%. This idea is elaborated upon in Section 4.3.

2.4.1.2 Heuristics and biases

There exist multiple bias- and heuristic-based explanations of the normative violations we have discussed in this article, including an associative reasoning bias (Rehder, 2014), the rich-get-richer bias (Rehder & Waldmann, 2017), the monotonicity assumption[7], conflict aversion, ambiguity aversion (Rottman & Hastie, 2016), and the Quantum Probability model (Trueblood et al., 2017). Except for the associative bias (Rehder, 2014), the authors of these explanations have not explicitly considered predictions related to time pressure as they present their theories as descriptive (in contrast with the Mutation Sampler, which is a process model). However, typically the reliance on heuristics and biases increases when people are under time pressure (Gigerenzer & Goldstein, 1996; Gigerenzer & Selten, 2002; D Kahneman & Frederick, 2002; Rieskamp & Hoffrage, 1999, 2008). Hence, if these proposed heuristics and biases function as typical heuristics and biases do (the authors provide no reasons for why we should not expect them to), we would expect that the increased reliance on them due to time pressure would result in larger normative violations. For example, Rehder and Waldmann conclude that people's inferences are "a product of an interaction between the normative model and the rich-get-richer principle" (Rehder & Waldmann, 2017, p. 255). Assuming that the richer-get-richer bias functions as a typical bias, we would expect the relative contribution of the richer-get-richer bias to increase. On the other hand, when there is little time pressure, participants would be able to engage in more deliberative strategies that would decrease the reliance on heurstics and biases, and thus decrease violations. Similarly, a dual-systems perspective that attributes the reasoning errors to System I responses (such as discussed in (Rehder, 2014)), would also wrongly predict an increase of these errors under time pressure as time pressure is known to increase intuitive responding (Evans, 2008; Evans & Curtis-Holmes, 2005).

While we do find an effect of time pressure on overall accuracy, we do not find a systematic effect on Markov violations and failures to explain away, which is not consistent with this perspective on the working of biases and heuristics under time pressure. We could speculate that some heuristics are more affected by time pressure than others. Heuristics and biases that explain Markov violations and failures to explain away by inducing correlations between variables (the associative bias (Rehder, 2014); the richer-get-richer principle (Rehder & Waldmann, 2017); the

[7] Associative reasoning, the richer-get-richer principle and the monotonicity assumption are functionally the same when considering the causal networks we have used in our experiments.

monotonicity assumption (Rottman & Hastie, 2016)) could be more resilient to time pressure. These heuristics can be implemented by the simple tallying of positive and negative cues in the stimulus, i.e. a tallying strategy (Gigerenzer & Gaissmaier, 2011), which could be such a fast and automatic strategy that it is not affected by time pressure. Rehder (2014) discusses the idea that associative reasoning might not be affected by more extensive deliberation, especially when reasoners are not confronted with a cue that they are wrong, as associative responses are so easy to generate. In addition, he raised the possibility that while being able to reason causally, people might often lack the metacognitive awareness that associative and causal reasoning might result in different responses. A possible explanation is then that the heuristics accounting for conservatism, like ambiguity and conflict aversion (Rottman & Hastie, 2016), might be more sensitive to time pressure as they are directly related to uncertainty which is affected by time pressure. However, these heuristics also partly predict Markov violations and failures to explain away. Hence it is unclear whether they are part of the right explanation, since if these heuristics are truly responsible for conservatism and are affected by time pressure, we should have observed a systematic impact of time pressure on Markov violations and failures to explain away.

2.4.2 Different classes of violations and implications for theories of causal reasoning

None of the theories just discussed seem to be consistent with all our observations. Nevertheless, our results do point to a way forward. Our results indicate that not all systematic non-normative reasoning patterns in causal reasoning are the result of a single cognitive process or mechanism.

The sensitivity of conservative inferences to time effects and their relationship to confidence suggest that they have a different source than Markov violations and failures to explain away. This seems probable considering that these errors are of different types. While Markov violations and failures to explain away refer to the not adhering to normative (in)dependence relationships between certain causal variables, conservative inferences are not related to such (in)dependencies stipulated by CBNs. Moreover, Markov violations and failures to explain away are relational in the sense that they require the comparison of multiple judgments, while conservatism is measured only in comparison with a normative response. In light of these theoretical considerations and our results it seems clear that we should view Markov violations and failures to explain away as belonging to a different class of errors than conservatism. This is at odds with existing theories of causal reasoning that attempt to explain all three errors partly with the same mechanism (the Mutation Sampler in (Davis & Rehder, 2020); Beta Inference, Conflict, and Ambiguity Aversion in (Rottman & Hastie, 2016)).

Our findings suggest that conservative inferences could be the outcome of a more general phenomenon related to uncertainty. Indeed, conservatism has been found in a wide variety of tasks in which participants have to judge probabilities (Costello & Watts, 2014; Erev et al., 1994; Hilbert, 2012; Phillips & Edwards, 1966; Zhu et al., 2020). When participants are uncertain, as evidenced by confidence judgments, they might use ‘default’ or safe response options, such as the middle of the scale (Kolvoort et al., 2021). This ties into a larger issue concerning the interpretation of probabilities (Fischhoff & de Bruin, 1999; Hájek, 2012). Simply put, a response of ‘50%’ can represent a strong belief of a participant that the correct answer is ‘50%’, but such a response can also represent the lack of a strong belief, i.e. epistemic uncertainty. That we found conservatism to be related to uncertainty provides evidence for the latter interpretation: participants seem to respond near 50% because they are uncertain about the correct answer, not because they necessarily believe the correct answer to be near 50%. Following this line of reasoning, we can view participants' probability responses as an expression of second-order probabilities, i.e. the probability that a probability (judgement) is correct or 'epistemic reliability' (Goldsmith & Sahlin, 1983). This interpretation of responses around 50% is bolstered by recent findings in non-probabilistic causal judgements tasks. In these tasks participants are asked to rate to what extent one factor caused another, and participants were found to use the middle of the scale when they were uncertain (O'Neill et al., 2022).

We conjecture that the conservatism we observe is due to participants using priors on the inference[8]. That is, participants could include prior knowledge about good responses to a query. When people are presented with a stimulus, they integrate the evidence they gain from the stimulus with prior information. If they are unable to gather much evidence for a response from a stimulus, e.g. in the case of a conflict or ambiguous trial, the prior will dominate. This explanation would be in line with recent trends modelling cognition using Bayesian principles (Knill & Pouget, 2004; Oaksford & Chater, 2020; Sanborn & Chater, 2016; Tenenbaum et al., 2011). Such a mechanism could explain conservatism overall, and additionally the effect of time pressure on conservatism; with more time pressure less evidence can be gained from reasoning based on the stimulus and hence the effect of the prior on the judgment increases. In the case of extreme uncertainty, when there would be no to little incorporation of information from the stimulus, this might result in responding at exactly 50%. This could explain the spikes of responses at 50% in causal judgement studies, if we assume that the prior knowledge that participants incorporate puts an emphasis on 50%. Our results are in line with viewing confidence judgements as an indication of the relative contribution of a prior to people's judgements.

2.5 CONCLUSION

Our study for the first time shows that causal reasoning mechanisms are systematically affected by both external (deadlines) and internal (passing of time) time pressure. This revealed a complex pattern of macro- and micro-SAT, which can be used to test and inform theories of causal reasoning. It seems that conservative inferences are the result of a different cognitive mechanism than that responsible for Markov violations and failures to explain away, as the former is related to time pressure and confidence while the latter are not. This study therefore also emphasizes the need for a wider range of behavioral phenomena than just plain mean responses to be incorporated into theories and computational models of causal reasoning. Incorporating more detailed

[8] Priors on causal parameters (in particular causal strengths) have been proposed before (Lu et al., 2008; Meder et al., 2009; Rottman & Hastie, 2016; S. Yeung & Griffiths, 2015). One problem with such an approach is that it can't explain differences in conservatism across inferences (Rottman & Hastie, 2016). Using an prior on the inference can explain such differences. Here we should interpret the stimulus itself as the likelihood (and not learning data or other information on the causal structure as with priors on causal parameters). Hence if the stimulus is clear-cut, providing consistent cues (as in e.g. P(Xj | Y=1, Xj=1), its influence will dominate the posterior (that is, the judgment).

phenomena – like the (in)sensitivity to time pressure, confidence (O’Neill et al., 2022), but also between- (Davis & Rehder, 2020; Rottman & Hastie, 2016) and within-participant (Kolvoort et al., 2021) variability – will lead to better theories. Other domains of judgment and decision-making have benefitted enormously from such a turn.

3 Variability in Causal Judgments

Abstract

People's causal judgments exhibit substantial variability, but the processes that lead to this variability are not currently understood. In this paper, we studied the within-participant variability of conditional probability judgments in common-cause networks by asking participants to respond to the same causal query multiple times. We establish that these judgments indeed exhibit substantial within-participant variability. This variability differs by inference type and is related to the extent to which participants commit Markov violations. The consistency and systematicity of this variability suggest that it may be an important source of evidence for the cognitive processes that lead to causal judgments. The systematic study of both within- and between-person variability broadens the scope of behavior that can be studied in causal cognition and promotes the evaluation of formal models of the underlying process. The data and methods provided in this paper provide tools to enable the further study of within-participant variability in causal judgment.

This chapter has been adapted from: Kolvoort, I.R., Davis, Z.J., van Maanen, L., & Rehder, B. (2021). Variability in Causal Judgments. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

3.1 INTRODUCTION

Causal relationships are a central way in which humans experience the world. Causal knowledge affects what decisions we make, how we categorize objects, and what counts as a good explanation (see Sloman, 2005; Sloman & Lagnado, 2015; Waldmann, 2017a). One of the main tools in studying causal cognition has been the theoretical framework known as Causal Bayesian Networks[9] (CBN; Pearl, 2009). CBNs have been shown to provide a generally good account of the causal judgments that people make. However, causal graphical models provide a computational level account that specifies what causal judgements are made, but not necessarily how people make them. Given the importance of causal knowledge to higher-level cognition, surprisingly little attention has been given to the processes by which people make such sophisticated judgments. In addition, recent empirical investigations have identified multiple systematic deviations from CBN predictions in human data (Davis & Rehder, 2017; Rehder, 2014; Rehder & Waldmann, 2017; Rottman & Hastie, 2016). To account for these deviations, researchers have developed multiple, mostly descriptive, theories (Rehder, 2014, 2018; Rottman & Hastie, 2016; Trueblood et al., 2017). These theories have been hard to distinguish as they have been developed to account for the same data, and they vary in how much light they shed on the process by which people generate causal judgments.

How can we evaluate which process generated a judgment? The predominant approach is to assess the predictions of multiple models against the average judgments of participants. This approach is principled and effective, but in a field as rich as causal cognition, utilizing only averaged data has not been able to convincingly identify the best model out of the multitude that have been proposed (Rehder, 2014, 2018; Rottman & Hastie, 2016). Other data can help with this underdetermination problem. For example, in judgment and decision making the popular diffusion decision model has exhibited considerable success in not merely accounting for mean judgments, but also explaining full distributions of response variables (Ratcliff et al., 2016). In this project, we use the full distribution of causal judgments as a new source of information about underlying cognitive processes involved.

A few studies have remarked on the considerable variability in human causal judgments (Davis & Rehder, 2020; Rehder, 2014; Rottman & Hastie, 2016). However, it is hitherto unclear to what extent that variability represents within- or between-participant variability. Measuring within-participant variability requires multiple independent measurements of the same type of judgment from the same participant. Whereas some studies have measured the same judgment type more than once, practical concerns have prevented them from being gathered on a scale large enough to estimate a full response distribution.

The major difficulty is that asking subjects to make the same judgment repeatedly is likely to yield measures that are not independent. Other areas that commonly elicit repeated measurements often have stimuli such as random-dot motion arrays that can be presented repeatedly without participants' awareness. Typical causal judgments do not have this property. Stimuli like ours that are composed of discrete symbols (such as states of causal variables) are susceptible to be recognized and memorized. This can be a problem particularly for studying higher-order

[9] Also known as Causal Graphical Models

cognition, such as causal reasoning, due to its more deliberative and conscious nature. In fact, storing previous judgments for future use has been proposed to be an important source of computational savings for limited agents (Dasgupta & Gershman, 2021). Our challenge was to design an experiment that elicits independent judgments for repeated causal queries by reducing the likelihood that participants' judgments are informed by prior computations or memory. We attempt to do so by using a symmetrical causal structure, querying participants regarding both the absence and presence of causal factors, and using the same parametrization across different domains in order to obtain multiple measures.

This project aims to understand some features of within-participant variability in causal judgments. Firstly, we aim to establish whether there is meaningful within-participant variability in causal judgments. Secondly, we look to compare variability across different inference types. Are there differences between forward (from cause to effect) and backward (from effect to cause) inferences? Does the information on which a participant is to base their inference impact variability? Thirdly, we investigate whether individual level variability is related to a tendency to commit an important systematic reasoning error known as Markov violations. We then describe potential models of variability in the causal reasoning process and provide a comparison of the observed variability against their qualitative predictions. We conclude by discussing the connections between the patterns of variation observed in our study with existing findings in causal cognition and opportunities for the use of full response distributions in the study of how people reason with causal information.

3.2 EXPERIMENT

3.2.1 Materials

We tested causal judgments in five domains: biology, astronomy, economics, meteorology, and sociology. Participants were first told that the domain they were about to study included three binary variables. For example, in the domain of economics they were told that interest rates could be either low or normal, trade deficits that were small or normal, and retirement savings that were high or normal.

Participants were then presented with a description of two causal relations that formed a common cause network in which one variable (henceforth referred to as Y) was a cause of the two others (X_1 and X_2). Each causal relationship was generative and included a description of the mechanism responsible for that relationship. An example in the domain of economics is "Low interest rates cause small trade deficits. The low cost of borrowing money leads businesses to invest in the latest manufacturing technologies, and the resulting low-cost products are exported around the world." All these materials have been validated by and used in multiple other studies (Rehder, 2014, 2018; Rehder & Waldmann, 2017).

3.2.2 Procedure

Subjects first studied several screens of information about the overall task that established the domains being studied and the types of inferences that would be presented during the study. Then, for each domain, initial screens presented a cover story and a description of the domain's three variables and subsequent screens presented the two causal links and a diagram of those links. A common cause network was used in every domain, and participants were informed that each variable's base rate was 50% and that each cause produced its effect "75% of the time".

When ready, participants were asked three multiple-choice questions to assess their understanding of the causal relationships. This comprehension check established that they had learned which variables were causally related, the direction of those relationships, and that the relationships were probabilistic rather than deterministic. Participants were given three attempts to answer all questions correctly. Once they answered all questions correctly or after the third attempt participants could continue with the experiment.

Subjects were then presented with the inference test. Each trial presented the values of one or two variables and asked to predict the state of another. For example, a subject might be told that an economy has low interest rates and a normal trade deficit and be asked the probability of it having a high level of retirement savings. Subjects entered their response by moving a tick on a rating scale whose ends were labeled 0% and 100%. As an attention check, participants were asked a comprehension check question at the end of each block. The order of the five domains, and the 24 test questions within each domain, was randomized for each participant.

3.2.3 Design and Participants

We chose six particular inference types to be tested based on the relevant comparisons they would allow. Firstly, we wanted to compare diagnostic or 'backward' inferences in which one has to judge the probability of a cause based on knowledge of its effects with predictive (or 'forward') inferences in which one reasons from cause to effect. Second, we assessed the effect of the information on which participants had to condition their inference: consistent information (where the states of the known variables are in line with the stipulated causal relationships, e.g. $X_i = 1$, $Y = 1$), inconsistent information (e.g. $X_i = 1$, $Y = 0$), and incomplete information (e.g. $X_i = 1$ and Y unknown). These factors lead to the six inference types presented in Table 3.1. To obtain multiple measurements, within each domain each inference type was queried four times by (a) varying whether the role of X_i was filled by X_1 or X_2 (possible because of the symmetry of the common cause structure) and (b) asking about both the presence and the absence of the to-be-inferred variable (using $P(X_i = 1|Y = 1) = 1 - P(X_i = 0|Y = 1)$). This resulted in each inference type being queried 20 times over the five domains and a total of 120 queries per participant. In those trials where we queried the absence of a variable, we flipped the responses around the midpoint of the probability scale. Table 3.1 also presents the normative conditional probabilities based on the 50% base rates and 75% causal strengths.

		Reasoning Direction	
		Predictive	Diagnostic
Information	Consistent	$P(X_i = 1 \| Y = 1, X_j = 1)$ = 80%	$P(Y = 1 \| X_i = 1, X_j = 1)$ = 94%
	Incomplete	$P(X_i = 1 \| Y = 1)$ = 80%	$P(Y = 1 \| X_i = 1)$ = 80%
	Inconsistent	$P(X_i = 1 \| Y = 1, X_j = 0)$ = 80%	$P(Y = 1 \| X_i = 1, X_j = 0)$ = 50%

Table 3.1 *The inference types tested in the experiment and their normative answers. Inference types varied with two factors, Reasoning direction (predictive or diagnostic) and Information (consistent, incomplete, or inconsistent) resulting in 6 inference types. Xs and Ys refer to variables, where the Xs are effects, and the Y is the cause in a three-variable common cause network (see* ***Figure 1.1****). The 1s and 0s refer to the presence or absence of an effect or cause.*

It is noteworthy that all the predictive inferences have the same normative probability of 80%. These inferences have been shown to exhibit "Markov violations", a pattern of responses in which, rather than adhering to the independence relations between variables stipulated by CBN theory, participants' responses are instead influenced by independent and hence irrelevant variables (Rehder, 2014; Rottman & Hastie, 2016). For these inferences, the value of one effect (X_i) should not provide information regarding the other effect (X_j) once the value of Y is known.

All participants made all judgments for all five domains. 37 participants were recruited from Prolific (www.prolific.co) and received £5.70 for on average 47 minutes (*SD* = 20.1) of participation. 8 (22%) participants were removed from analyses for failing at least two attention checks, as had been established by the authors before the running of the study.

3.3 Results

As described in the Design section, our materials utilized multiple sources of redundancy to maximize the number of observations of a single inference. Results were collapsed over these factors for a total of 20 judgments per inference types per participant[10]. Figure 3.1 plots the individual response distributions per inference type. This plot shows substantial between-participant variability, as we see that some participants' responses are more spread out than others, and some participants exhibit bimodality in some or most judgements whereas others do not at all. We see similar patterns in the within-participant variability. The first thing to note is that there is substantial variability in each participant's responses. Moreover, the overall spread and the modality of the response distributions differs per inference type for many participants.

[10] Due to an error in the materials three diagnostic trials were removed from the Economics domain for all participants, resulting in 19 judgments for the diagnostic inferences.

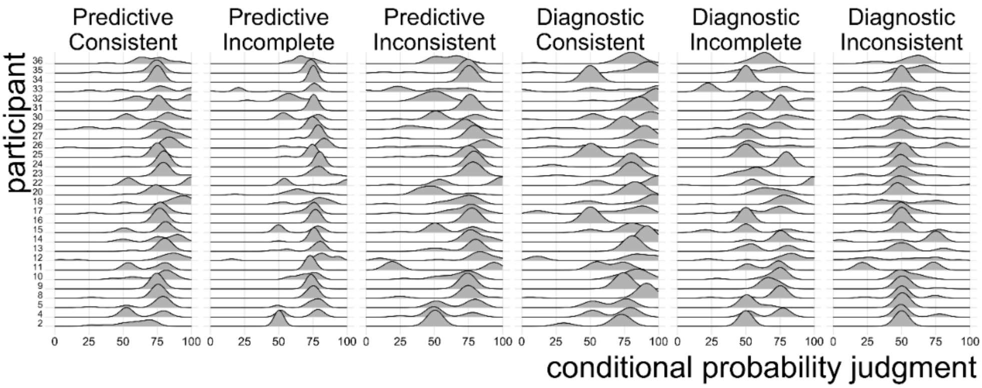


Figure 3.1 *Per participant distributions of responses for each inference type. Rows correspond to participants, columns correspond to judgment types, the x-axis indicates the responses in percentage points, and the height corresponds to kernel density estimate of participant responding at this probability. Each density plot is based on 20 responses.*

The response distributions per inference type averaged over all participants is illustrated in Figure 3.2. The first aspect to note is that the distributions vary by judgment type. If the only source of variability is unrelated to the process by which causal judgments are generated (such as general response noise), we would expect similar variability across judgments. The bimodality of the response distributions in Figure 3.2 is also noteworthy. In particular, we observe a “spike” of responses at 50%, which has been reported previously (Rottman & Hastie, 2016). This peak at 50% seems to vary along the Information factor, with the largest peaks for inconsistent inferences and smallest for inferences with consistent information. As expected, the peak is largest for inconsistent diagnostic inferences for which the normative answer is 50%.

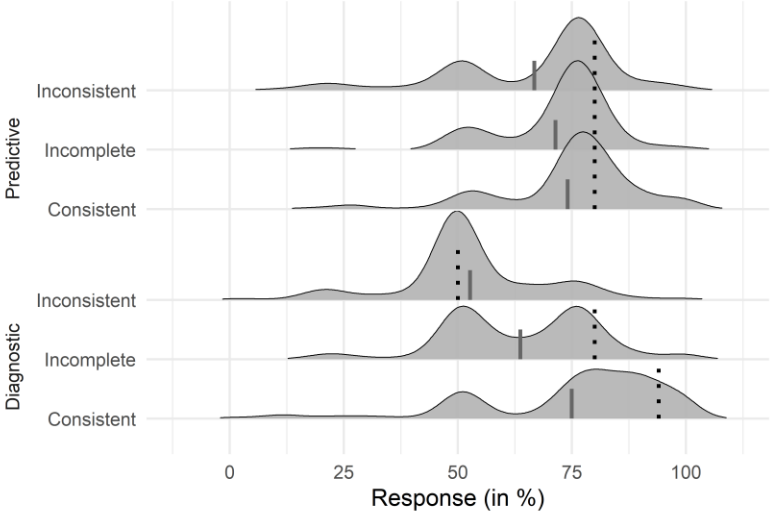


Figure 3.2 *Overall response distributions per inference type. Vertical grey lines indicate mean responses. Dotted vertical black lines indicate normative response.*

Figure 3.3 shows the means of within-participant standard deviations and mean judgments per inference type. Note in Figure 3.3 that while variability differs by inference type, it does not track with the mean, suggesting that these results are not driven by an artifact of the scoring system. We tested whether variability differs over the inference types using a repeated measures ANOVA with the standard deviation in responses as the dependent variable and Diagnostic (yes, no) and Information (consistent, incomplete, inconsistent) as factors. The main effect of Information is significant (*F*(2,140) = 9.58, *p* < .001, *BF* > 100). This indicates that the variability is lower for inferences with incomplete information (*Mean* = 10.4, *SE* = 1.4), than for inferences with complete information (*consistentMean* = 14.1, *SE* = 1.4, *inconsistentMean* = 13.7, *SE* = 1.4). We find mixed evidence of an effect of Diagnostic (*F*(1,140) = 4.24, *p* = .041, *BF* = .893). Variability was marginally higher for diagnostic inferences (*Mean* = 13.5, *SE* = 1.3) than for predictive inferences (*Mean* = 11.9, *SE* = 1.3) when conducting a post-hoc contrast (*difference* = −1.58, *SE* = 0.766, *t*(140) = −2.01, *p* = .041). There was no evidence for a Diagnostic × Information interaction (*F*(2,140) = 2.52, *p* = .084, *BF* = 1.16). That there are differences in variability over inference types suggest that it results from some underlying process of generating causal judgements.

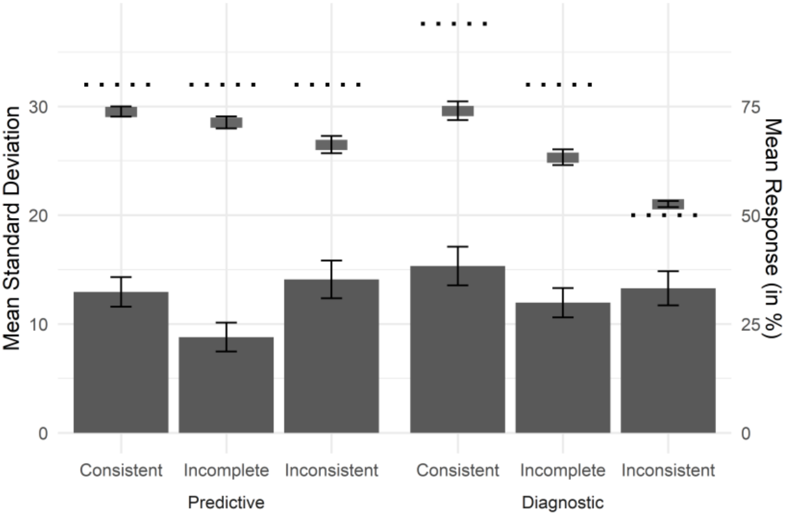


Figure 3.3 *Barplot: Mean within-participant standard deviations per inference type. Floating dashes: Mean responses per inference type. Black vertical lines indicate standard error. Horizontal dotted lines indicate normative probability.*

To test whether variability and Markov violations are related, we first separated participants into three (low, medium, high) equally sized groups based on the standard deviation of their responses on predictive inferences. Figure 3.4 plots the mean predictive judgments by variability group, revealing an apparent increase in non-normative responding as variability increases. We conducted an ANOVA on the responses on predictive inferences with Information as factor and participant's standard deviation as a covariate. We find a significant main effect of Information (*F*(2,1575) = 29.6, *p* < .001, *BF* > 100), which indicates that overall participants committed Markov violations, as normatively the Information factor should not have an effect as the normative response to all predictive inferences is 80%. We find evidence for a main effect of each participant's standard deviation (*F*(1,1575) = 51.2, *p* < .001, *BF* > 100), indicating that participants with more variable judgments overall give lower responses, this is also seen in Figure 3.4. Most interestingly, we find very strong evidence for an interaction between Information and the grouping variable (*F*(2,1575) = 12.5, *p* < .001, *BF* > 100), indicating that high variability participants commit larger Markov violations. This interaction is illustrated in Figure 3.4 by the thick black line, which becomes steeper (larger Markov violations) for the higher variability groups.

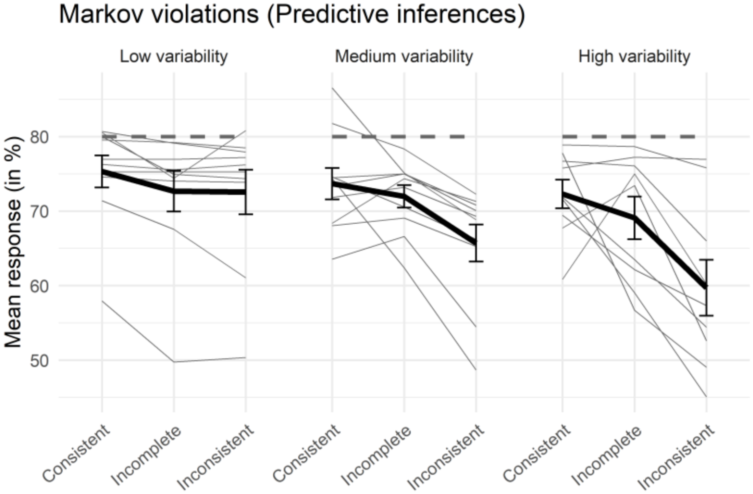


***Figure 3.4** Plots of Markov violations per variability group. Participants were first separated into three (low, medium, high) equally sized groups based on the variability in their responses on predictive inferences. Grey thin lines represent the mean responses of individual participants on the predictive inferences. Black thick lines represent the mean responses per group, the vertical bars indicate standard errors. The dashed lines represent the normative response of 80%.*

We also asked whether the observed variability was related to cross-domain variability or fatigue effects, rather than the reasoning process itself. We conducted a repeated measures ANOVA on the within-participant standard deviation using the order of blocks as presented as a predictor. We find an significant effect of block order (*F*(4,112) = 3.62, *p* < .001, *BF* = 4.15). Post-hoc contrasts reveal that the first block is significantly different from the latter blocks, which do not differ from each other (*Mean SDs*: first block 18.5, second 16.3, third 15.6, fourth 15.4, fifth 14.6). That variability stayed constant after the first block suggests that it is unlikely to be due to fatigue. This result also argues against strategy changes over the blocks, which indicates we largely succeeded in eliciting independent repeated judgements. One would expect an increase in variability over the latter blocks had subjects recognized that they were repeatedly being asked the same judgment type and so settled on a consistent response strategy. To test whether the content domains affected variability we conducted a repeated measure ANOVA on the within-participant standard deviation with Inference type and Domain as factors. We find evidence against an effect of Domain on variability (*F*(4,789) = 1.03, *p* = .39, *BF* < .01) and against an interaction of Domain with Inference type (*F*(20,789) = 1.01, *p* = .44, *BF* < .01).

3.4 SOURCES OF VARIABILITY

What processes explain the variability in responses to causal queries? As a guide for future research, in this section we outline a number of candidate models of the variability in conditional probability judgments. While fitting these models against participants' response distributions is beyond the scope of this paper, we discuss the correspondence of their qualitative predictions with the results of our experiment.

One possibility is that the observed variability in responses is entirely independent from the cognitive process by which a causal judgment is generated. It could be that people have a stable causal representation and strategy to arrive at a causal judgment, but that the process of responding to a query results in some noise, e.g. through motor noise in using a slider or some general task noise. In this case one would expect response distributions that are centered at the normative answer, such as predicted by the Beta inference model (Rottman & Hastie, 2016). Our findings provide evidence against this possibility: response distributions are often multi-modal (see Figure 3.2), and variability differs by inference type and seems to be related to patterns of non-normative responding.

Another possibility could be that the source of variability in causal judgments stems from uncertainty about the parameters of the described causal network. For example, rather than believing that the causal strength of A on B is precisely .75, this value may have some variance. Because the CBN framework models causal judgments as being computed from a causal network, this would result in variation in the resultant causal judgments if a participant reasoned with slightly different parameter values for each inference. Such an account may explain increased variability in diagnostic inferences, as according to the CBN framework these require the processing of an additional parameter, the base rate of the cause (Fernbach et al., 2011). It is unclear how this approach would explain why judgments where two pieces of information are given are more variable than only one piece, as the CBN framework would predict that there is no change in the number of parameters that need to be considered. In addition, this CBN-based account is incompatible with our observed Markov violations. See the Discussion section for further discussion of these patterns of judgments.

One salient pattern in the data is the "spiking" at 50%. This has also been observed in between-subjects data like that from Rottman and Hastie (2016). Responses at 50% may reflect guessing or responding in some default manner. One possibility is that one of the above models, in combination with a probability of responding at 50%, can explain the observed variability. While this may account for some variance, such a model would still need to explain why the prevalence of these 50% responses in varies by inference type. In particular, it has to provide an account of why those spikes are largest for inconsistent inferences and smallest for consistent inferences. One explanation could be that participants are more likely to guess when the information provided for an inference is more ambiguous.

Both response noise and uncertainty about parameters are compatible with the normative CBN framework being the underlying process used to generate causal judgments. Other models of causal reasoning predict variability as a consequence of the reasoning process itself. The mental model theory of causation stipulates that causal judgments are rendered from imagined concrete states, as determined by the causal structure that is being reasoned about (Johnson-Laird &

Khemlani, 2017). A similar account from Davis & Rehder (2020) models these imagined states as being the result of a structured mental search through the space of possible situations, in the form of a Markov Chain Monte Carlo sampling process. The stochastic nature of this sampling process introduces variability. And while not explicitly designed as a process model, quantum models of causal reasoning may make unique predictions by virtue of participants varying in the dimensionality of their representations (Trueblood et al., 2017).

While all of these accounts make predictions about response distributions, the Mutation Sampler is the only model for which predictions about response distributions have been explicitly reported (Davis & Rehder, 2020). One of these predictions is that of spikes at 50% (resulting in multimodal distributions), which appear to be borne out in our data. Moreover, the Mutation Sampler predicts an increase in spikes for inconsistent trials because it incorporates a mechanism for default responding at 50% when the sampling process does not provide information to answer the query. This is more likely for inferences with inconsistent information as states with incongruous variable values are sampled less often.

3.5 DISCUSSION

This article takes the first step in bringing the field of causal reasoning in line with other domains of cognitive science that take into account the variability of judgments and not just their averages. As exemplified by the prolific use of the diffusion decision model (Ratcliff et al., 2016), response distributions provide more sensitive signals to underlying cognitive processes. We consider the development of an experimental design that elicits multiple measurements of the same causal query to be a primary contribution of this project.

Our findings show, for the first time, that there is indeed meaningful within-participant variability in causal reasoning. That our data exhibit similar variability to that in between-subjects studies, suggests that it largely arises from the processes by which individuals generate causal inferences. That it varies with the type of causal inference supports the additional conclusion that the variability at least partly reflects a decision-making process rather than noise (e.g., noise in motor responses) or some other factor about individual participants (Rottman & Hastie, 2016).

We found mixed evidence that the direction of reasoning might matter: Diagnostic (from effect to cause) inferences were overall more variable than predictive (cause to effect) inferences. This squares nicely with the often-repeated claim is that it is easier to think in the direction from cause to effect (Tversky & Kahneman, 1982). This finding adds to the existing empirical literature on differences between diagnostic and predictive reasoning, which has reported that people take longer to respond to diagnostic queries (Fernbach & Darlow, 2010) and that they do not neglect possible alternative causes (which they tend to do for predictive inferences; Fernbach et al., 2010). It has also been argued that diagnostic reasoning is more comparative (Fernbach et al., 2011), and CBN theory stipulates that diagnostic reasoning requires the incorporation of additional information, namely the prior probability of the cause. That the more variable diagnostic judgments have also been found to be more difficult suggests that the observed response distributions reflect the processes by which these judgments are rendered.

The information provided to participants in conditional inferences also matters: knowledge of all non-queried variables leads to an increase in variability, while incomplete information seems to reduce it. These findings are somewhat surprising. One might expect that additional information would result in less uncertainty over the possible values of an unknown variable. We find the opposite. It might be that more pieces of information result in more variability by virtue of there being more ways to process two pieces of information versus one. A related explanation appeals to stimulus encoding. When more pieces of information are provided as part of the stimulus, it might be more probable that there is more variation in whether one or more pieces of the stimulus are encoded incorrectly on a portion of the trials.

We also found a relationship between violations of the causal Markov condition and variability over participants. Participants who are more variable tended to exhibit stronger Markov violations. This finding squares with a large literature suggesting that Markov violations are a key source of evidence for the claim that the normative CBN framework is not an accurate model of the true underlying process that people use to draw causal judgments (Rehder, 2014; Rottman & Hastie, 2016; Trueblood et al., 2017). Importantly, Markov violations are by definition incompatible with any model that uses the CBN framework as its core representation, and therefore defies simple interpretations of the observed variability as response noise or uncertainty about the parameters of the causal model. Instead, it appears to signal that a common underlying process drives both Markov violations and part of the observed variability. This underlying process may be related to individual factors. One such factor might be a difference in reasoning strategy or style, which would be in line with findings relating Markov violations to tendency to engage less in reflective thought (Trueblood et al., 2017). Another possible factor may be limitations in working memory capacity, as proposed by Davis and Rehder (2020).

The experimental design used in this study has limitations. A major experimental obstacle was eliciting 24 unique judgments for identical causal queries. Variability in judgments may have resulted from variability in interpretation of experimental materials, rather than in the causal reasoning process itself. For example, people may have different beliefs about the causal relationships between societal factors than between features of stars. We believe this possibility cannot account for all the observed variability, as we found no differences in variability over domains and our usage of the same parameters for all domains reduces this possibility further (see also earlier discussion of uncertainty in parameters as a source of variability). Another limitation is our use standard deviation as an index of variability. Since the distributions are not unimodal this measure does not necessarily capture all relevant information in the response distributions. Lastly, we only tested a subset of the possible inferences in one particular causal inference task. The extent to which our findings apply to other inferences or tasks is an open question.

We discussed the correspondence between our findings and the qualitative patterns of variability in potential models of the causal reasoning process. Fitting full response distributions is a challenging computational and statistical problem that goes beyond the scope of this paper. We do wish to emphasize that future efforts should focus on this challenge, as modeling more than just averaged judgments will help improve our understanding of the cognitive processes underlying causal reasoning.

3.6 Conclusion

Causal reasoning is a core cognitive activity. Understanding the processes by which people generate causal judgments will help us better understand a range of cognitive activities from decision-making to categorization. In this paper we presented the first investigation of within-participant variability in causal judgments. This variability differs by inference type, is related to systematic reasoning errors, and is not easily explained by simple additions to the dominant CBN framework for causal inference. We hope that the data and methods presented in this paper will be useful in broadening the scope of behavioral signals used to study how people draw causal inferences.

Part 2

Computational cognitive modeling of causal reasoning

4 THE BAYESIAN MUTATION SAMPLER EXPLAINS DISTRIBUTIONS OF CAUSAL JUDGMENTS

Abstract

One consistent finding in the causal reasoning literature is that causal judgments are rather variable. In particular, distributions of probabilistic causal judgments tend not to be normal and are often not centered on the normative response. As an explanation for these response distributions, we propose that people engage in 'mutation sampling' when confronted with a causal query and integrate this information with prior information about that query. The Mutation Sampler model (Davis & Rehder, 2020) posits that we approximate probabilities using a sampling process, explaining the average responses of participants on a wide variety of tasks. Careful analysis, however, shows that its predicted response distributions do not match empirical distributions. We develop the Bayesian Mutation Sampler (BMS) which extends the original model by incorporating the use of generic prior distributions. We fit the BMS to experimental data and find that, in addition to average responses, the BMS explains multiple distributional phenomena including the moderate conservatism of the bulk of responses, the lack of extreme responses, and spikes of responses at 50%.

This chapter has been adapted from: Kolvoort, I.R., Temme, N., & van Maanen, L. (In press). The Bayesian Mutation Sampler Explains Distributions of Causal Judgments. *Open Mind*. Preprint DOI: 10.31234/osf.io/9kzb4.

4.1 INTRODUCTION

Causal reasoning is a core facet of human cognition. Dealing with causal relationships in the world and using them to our advantage is a crucial part of our abilities. Causal cognition ties into most (if not all) judgments and decisions (e.g. Hagmayer & Osman, 2012; Rottman & Hastie, 2014). Most of what we do and think is at least partly based on the perception of and reasoning about causes and effects in the world. This makes it an important aim in cognitive science to understand how we think and reason about causes and effects.

The current work addresses probabilistic causal reasoning. An example of this is when someone tries to judge the probability that they will be late for work after hearing on the radio that there has been a traffic accident nearby. To make such a judgment one needs to use their knowledge of a causal system. In this case such knowledge could be represented as X→Y→Z, where X stands for a traffic accident, Y for a traffic jam, and Z for being late for work. In a typical probabilistic causal reasoning experiment people are first taught about a particular causal system (for example X→Y→Z), after which they are asked to make certain inferences, i.e. compute certain (conditional) probabilities. An example of such an inference is "what is the probability of Z (being late for work) given that X (a traffic accident) happened?". These experiments provide a window into how participants make use of causal information to come to a specific judgment. The current work will focus on using cognitive modelling to understand how people make such judgments using their causal knowledge.

Over the last decades so-called causal Bayesian networks[11] (CBNs; Pearl, 2009; Spirtes et al., 2000) have achieved remarkable success in modeling human behavior across a variety of tasks related to causal learning, categorization, reasoning and inference (e.g. Ali et al., 2011; Bramley et al., 2015; Cheng, 1997; Coenen et al., 2015; Fernbach & Erb, 2013; Griffiths & Tenenbaum, 2005, 2009; Hagmayer, 2016; Hayes et al., 2014; Holyoak et al., 2010; Krynski & Tenenbaum, 2007; H. S. Lee & Holyoak, 2008; Lu et al., 2008; Meder et al., 2014; Rehder, 2014; Rehder & Burnett, 2005; Shafto et al., 2008; Steyvers et al., 2003; Waldmann & Hagmayer, 2006). These models provide a concise representation of causal systems and a particular formal logic that specifies how one can learn, draw inferences, and update causal knowledge based on interventions in the system.

CBNs, as models of human cognition, are often understood as explanations on the computational level (Marr, 1982). That is, they provide an account of what problem needs to be solved, but not how to solve them. This is because formal computations with CBN models tend to be computationally expensive and so are thought not be feasible as a way for us humans to solve problems. Instead of directly doing these Bayesian computations, recent work in the cognitive sciences has argued that people use sampling to solve such computationally intensive problems in a variety of domains (Bonawitz, Denison, Griffiths, et al., 2014; Dasgupta et al., 2017; Hertwig & Pleskac, 2010; Lieder et al., 2012; Vul et al., 2014; Zhu et al., 2020). This *sampling approach to cognition* proposes that we solve problems by way of first drawing samples, either from memory or an internal generative model, and then generating judgments based on the

[11] Also known as causal graphical models, graphical probabilistic models, or causal Bayes' nets.

information in these samples. In this way we can reason about probabilities without the need to explicitly represent probabilities. Often the sampling approach is modelled using Markov chain Monte Carlo (MCMC) processes (see Dasgupta et al., 2017). Davis and Rehder (2020) applied this sampling approach to the domain of causal cognition, developing a model that samples over possible states of a CBN to make causal judgments. This so-called Mutation Sampler model (MS) provides an algorithmic level explanation (Marr, 1982) as it describes *how* humans reason causally. It proposes a sampling mechanism for how we generate causal judgments and has been successful in explaining average responses on a variety of tasks (Davis & Rehder, 2020).

However, while accounts of average responses abound, the common observation of substantial variability in causal judgments has received less attention and is often left unexplained (Davis & Rehder, 2020; Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rehder, 2014, 2018; Rottman & Hastie, 2016). This is an unfortunate gap in the literature, as variability in behavior can be informative of the cognitive mechanisms involved and so can help constrain the development of theories (e.g., as has been done in the domain of decision-making; Ratcliff, 1978). In this paper, we will analyze the distributional predictions of the MS and ultimately extend it with the incorporation of priors to provide an explanation of some of the observed variability in causal judgments.

4.1.1 Sampling theory and the Mutation Sampler

The MS is a sampling model of causal reasoning that accounts for many observed behavioral phenomena including deviations from the normative CBN model (Davis & Rehder, 2017, 2020; Rehder & Davis, 2021). Before discussing the model in more detail, it is important to note that the original authors argue for four psychological principles to govern causal reasoning and that the MS is but one formalization of these principles. We will refer to these principles in this section, for a more detailed discussion on these psychological claims and the exact formalization of the MS we refer to the original paper (Davis & Rehder, 2020).

The MS proposes that people engage in sampling over states of a causal network to make inferences and as such describes the process by which people generate causal judgments. This proposal is built on the psychological claim that people reason about concrete cases and not about abstract probabilities (Principle 1). The concrete cases here are causal networks instantiated with particular values, i.e. they are particular casual network states (see Figure 4.1 for three-variable causal networks). These concrete cases or causal network states are obtained from memory or through simulations using an internal generative model. Subsequently these samples are used to compute inferences based on the relative frequencies of certain events in the chain of generated samples.

The chain of samples generated by this scheme converges to reflect the actual normative joint distribution when the number of samples becomes sufficiently large. This means that judgments based on a large number of samples approximate the true probabilities. However, people do not respond normatively and nor does the MS. Two factors contribute to the non-normativity of judgments based on the mutation sampling process: (1) the starting point of the process is biased (to prototypical states) and (2) the number of samples (or ‘chain length') is limited.

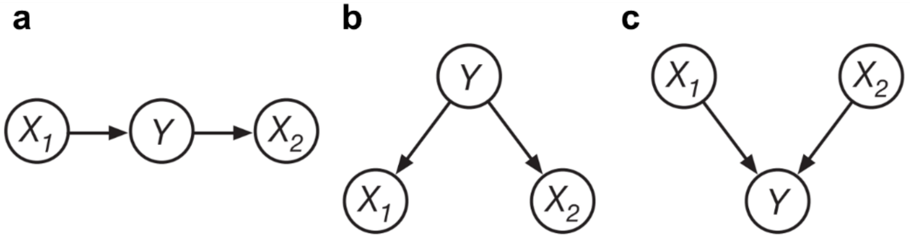


Figure 4.1 *Three-variable causal networks. The circles represent causal variables, the arrows represent causal relationships. Throughout this manuscript we will use Y to refer to the middle variable, and X_i to refer to a terminal variable. a. chain structure, b. common cause structure, c. common effect structure*

Both these factors, limited sampling and biased starting points, lead to a probability distribution that overestimates the likelihood of states where more variables have the same value. That is, it is biased towards prototypical states (see Figure 3 in Davis & Rehder, 2020). Since a chain of samples always starts at a prototypical state and has limited opportunity to reach states that are very different from these prototypical states (due to limited number of samples), the predicted probability distribution places more probability density on states based on their closeness to the prototypical states. This effect is stronger when the number of samples taken is small.

Davis and Rehder (2020) provide psychological justifications for these aspects of their model that lead to non-normative responding. With regard to the biased starting points, Davis and Rehder suggest that “prototypes readily come to mind as plausible states at which to start sampling because, if one ignores the details of the causal graph such as the strength, direction, or functional form of the causal relations, they are likely to be viewed as having relatively high joint probability” (Principle 3; Davis & Rehder, 2020, pp. 6). Assuming generative causal links, this is the case because a prototypical state is always consistent with the causal relationships in that there are no cases in which effects are absent while their causes are present and vice versa. Moreover, these prototype states are generally high probability states and so are good starting points for convergence. Taken together, when we start thinking about a causal system we start in a simple state, that is likely to be remembered or generated as it occurs often and is consistent with all causal relationships in the network.

The second aspect of the MS that leads to non-normative responding is that the chains of samples are of limited length (Principle 4). Other work on sampling approaches to cognition has shown that using only a limited number of samples can be rational when taking into account costs associated with taking more samples (see Dasgupta et al., 2017; Hertwig & Pleskac, 2010; Vul et al., 2014) As such the MS can be viewed as part of the new resource-rational analysis modeling paradigm in which the rational use of limited resources is a guiding principle (Lieder & Griffiths, 2020).

Davis and Rehder (Davis & Rehder, 2020; Rehder & Davis, 2021) fitted the MS to data from a variety of experiments concerning causal reasoning, categorization, and intervention. They found that the model fits better to participant responses than standard CBNs. Moreover, the model is able to account for multiple systematic reasoning errors, i.e. deviations from the normative CBN model, by virtue of the limited sampling and biased starting point mechanisms. For instance, the distorted joint distribution that the MS produces is able to account for Markov violations and failures to explain away, two hallmark behavioral phenomena in causal reasoning (Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Park & Sloman, 2013; Rehder, 2018; Rehder & Burnett, 2005; Rottman & Hastie, 2014, 2016). In addition to using existing data, Davis and Rehder (2020) ran an experiment that presented participants with causal graphs and asked them to generate data that they thought would be consistent with the causal structure. The data participants generated matched the distorted joint distributions produced by the MS. Taken together, the tested predictions of the MS seem to be in very good accord with experimental data.

However, not all aspects of the of predictions have been scrutinized. As the MS posits that the transitions from one state to another in the generation of a chain of samples are stochastic, it predicts not just a mean response (like the CBN estimate), but a full distribution of responses to an inference. This is in line with the MS being a process-model, since it models the process by which people generate causal judgments it should also produce the variability in responses seen on a variety of tasks (Davis & Rehder, 2020). However, following the literature on causal cognition at large, these distributional predictions have not yet received proper attention. The aim of the current paper is to assess these distributional predictions and use the distributional phenomena in empirical data to guide further development of the MS.

The rest of this paper is structured in two main parts. In the first part we analyze the distributions predicted by the MS and find that it cannot explain certain distributional phenomena observed in causal reasoning tasks. In addition, we provide theoretical arguments against the (resource-)rationality of the model, leading us to extend the model using priors in the second part. In the second part we introduce the Bayesian Mutation Sampler and test its distributional predictions. We conclude with a general discussion concerning both the theoretical and empirical advances made in this paper.

4.2 Analyzing the MS and Distributional Phenomena

The nature of the MS as a process-model makes it a useful tool to assess distributional phenomena in addition to the mean phenomena that have been extensively studied (see Kolvoort et al., 2021). It has been observed multiple times that causal judgments vary substantially (Davis & Rehder, 2020; Kolvoort et al., 2021; Rehder, 2014, 2018; Rottman & Hastie, 2016). The MS has not been used to study distributional properties of responses. The authors did present a figure indicating a qualitative similarity between the variability of responses in experiments 1A and 1B by Rottman and Hastie (2016) and the variability of the predicted responses by the MS with a mean chain length of 36 (their Figure 9; 2020). However, this value for the chain length is far from the average best fitting parameter (which was 12.7) that the authors found for a range of causal reasoning experiments (Davis & Rehder, 2020). Therefore, many questions remain regarding the predicted distributions of the MS. Here, we aim to assess these predictions under a range of different chain lengths.

To this end we simulated responses using the MS with multiple chain lengths using the same causal parameters[12] as in experiment 1A by Rottman and Hastie. We chose these parameters since the MS was fitted to that experiment originally and they are intended to theoretically neutral (Rottman & Hastie, 2016). Figure 4.2 present the results of the simulations together with empirical data from a recent causal reasoning experiment (Kolvoort, Fisher, et al., 2023). These data are causal probabilistic judgments, where participants had to judge the probability of a causal variable being present conditional on information about other causal variables in the network (e.g. $P(X_1 = 1|Y = 1)$). The experiment is described in more detail in a later section, for now it suffices to note that it used similar methods, including the exact same causal parameters, as experiment 1A by Rottman and Hastie (2016).

[12] The chain and common cause structure had the same causal parameters, with base rates of .5 for all variables. The effects in the network had a probability of 75% when their parent was present and 25% when it was not. In the common effect structure, the two causes combined by way of a Noisy-OR gate (Cheng, 1997) with causal strengths of 50% and with base rates of 50%. This meant that the effect had a 0% probability if no causes were present, 50% when one cause was present, and 75% when both causes were present (hence the base rate was .43 for the effect).

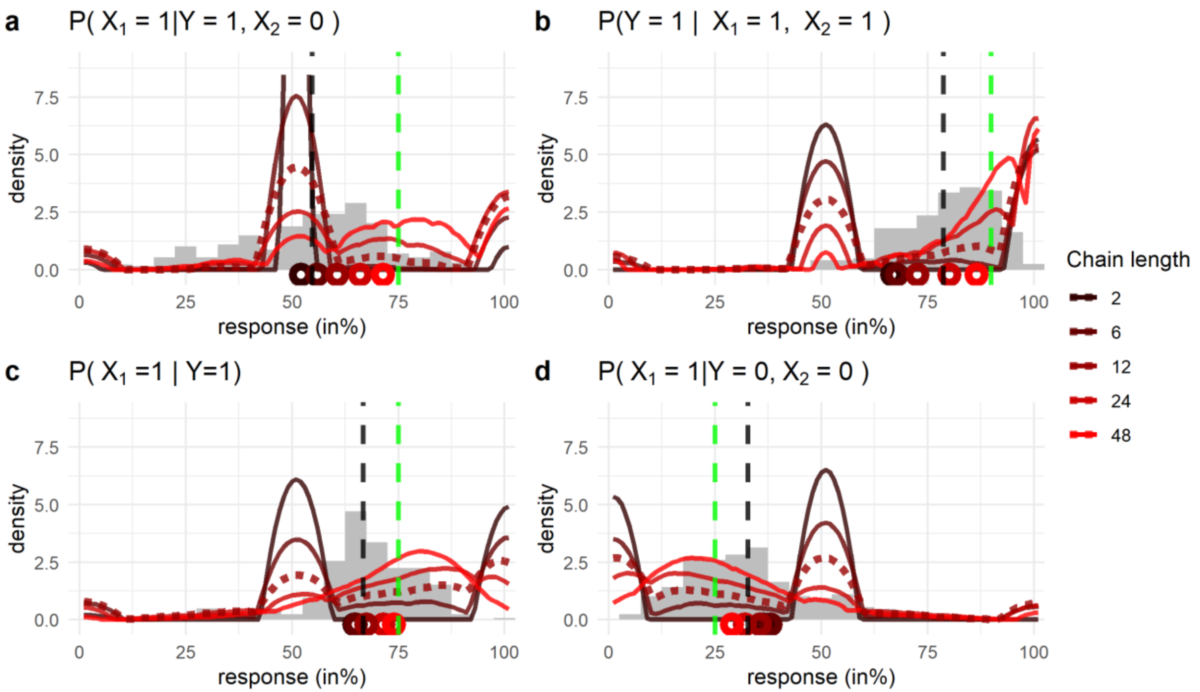


Figure 4.2 Predictions of the Mutation sampler and data from Kolvoort et al. (2023) on four different inferences (A: $P(X_1 = 1|Y = 1,\ X_2 = 0)$, B: $P(Y = 1|X_1 = 1,\ X_2 = 1)$, C: $P(X_1 = 1|Y = 1)$, D: $P(X_1 = 1|Y = 0,\ X_2 = 0)$). Plots A-D: The inferences were about a common cause structure where X_1 and X_2 refer to the two effects, and Y refers to the common cause. A state of 1 (e.g. Y=1) indicated a variable was present, a state of 0 indicated the variable was absent. Grey histograms are the participant responses. The red lines indicate predictions of the MS with different chain lengths. The dashed red line are the predictions with chain length of 12, close to the mean found by Davis and Rehder (2020). The predicted response distributions were generated by simulating 10,000 responses with the MS and smoothing the result using kernel density estimation. Vertical dashed black lines indicate mean participant response. Dashed green lines indicate the normative answer. The colored circles on the x-axis indicate mean predictions.

4.2.1 Mutation Sampler predicts extreme responses and no ‘moderate’ conservatism

Let us take the predictions of the MS with chain length 12 (dashed red lines in Figure 4.2) as a starting point for our discussion, since 12 is close to the mean chain length found for causal reasoning tasks (Davis & Rehder, 2020). What immediately stands out in Figure 4.2 is that the MS with a chain length of 12 predicts three peaks of responses for each inference at 0%, 50% and 100%. Spikes of responses at 50% have been reported in the literature on causal judgments (Kolvoort et al., 2021; Rottman & Hastie, 2016). To the contrary, the peaks at 0% and 100% seem not to correspond to empirical data. In fact, it is known that participants behave conservatively and tend to avoid the extremes of the scale (Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). This makes the predictions at 0% and 100% rather surprising.

To understand these predictions we have to take a closer look at the mechanisms causing these peaks. Firstly, the peak at 50%. These peaks are due to the MS defaulting to a 50% response for conditional probability queries when the causal network states required for the calculation are not reached at any point by the stochastic sampling process. Throughout this manuscript we will use Y to refer to middle variable, and X_i to refer to a terminal variable (see Figure 4.1). Let us say the required inference is $P(X_1 = 1|Y = 1,\ X_2 = 0)$ as in Figure 4.2a. In this case the sampler needs to visit the states $[X_1 = 1, Y = 1,\ X_2 = 0]$ and $[X_1 = 0, Y = 1,\ X_2 = 0]$ to compute the inference based on the relative frequency of these states in the chain. This computation is done using the (Kolmogorov) definition of conditional probability:

$$P\big(X_1 = 1\,\big|\,Y = 1, X_2 = 0\big) = \frac{P(X_1 = 1, Y = 1, X_2 = 0)}{P(X_1 = 1, Y = 1, X_2 = 0)\ +\ P(X_1 = 0, Y = 1, X_2 = 0)}$$

From which we can get an estimate for the conditional probability based on sample frequencies:

$$\hat{P}\big(X_1 = 1\,\big|\,Y = 1, X_2 = 0\big) = \frac{N(X_1 = 1, Y = 1, X_2 = 0)}{N(X_1 = 1, Y = 1, X_2 = 0)\ +\ N(X_1 = 0, Y = 1, X_2 = 0)} \qquad (EQ1)$$

Where N stands for the number of samples of that causal state. Now if the required states on the right-hand side of EQ1 are not visited their frequencies (or probabilities) are zero. In this case EQ1 would reduce to $\frac{0}{0}$ which cannot be computed and instead the MS defaults to 50%[13].

[13] Davis & Rehder (2020) implemented this mechanism by initializing the number of visits to each network state with 10^{-10}. When the required states for an inference are not visited EQ1 then simplifies to $10^{-10}/[10^{-10} + 10^{-10}] = .5$.

The predicted peaks at 0% and 100% come about similarly as the peak at 50%, however in this case only one of the two required states is not visited. Let us again consider the inference $P(X_1 = 1|Y = 1,\ X_2 = 0)$, which requires the state A: $[X_1 = 1, Y = 1,\ X_2 = 0]$ and B: $[X_1 = 0, Y = 1,\ X_2 = 0]$ to be visited by the sampler. In the case where state A is not visited, EQ1 simplifies to $\frac{0}{0\,+\,P(X_1=0,Y=1,X_2=0)}$ and we get a response at 0%. When state B is not visited by the sampler, EQ1 simplifies to $\frac{P(X_1=1,Y=1,X_2=0)}{P(X_1=1,Y=1,X_2=0)\,+0}$ and we get a predicted response at 100%.

To gain insight in how often the MS generates ‘default’ responses at 0%, 50%, or 100% we can estimate how often a particular network state is expected to be visited. We do this by simulating 10,000 chains of samples. Let us again regard $P(X_1 = 1|Y = 1,\ X_2 = 0)$ (Figure 4.2a), which requires visits to the states A: $[X_1 = 1, Y = 1,\ X_2 = 0]$ and B: $[X_1 = 0, Y = 1,$ $X_2 = 0]$ to be computed (see EQ1). We find that with a chain length of 12, the proportion of trials on which state A is not visited by the sampler is 0.49, the proportion where state B is not visited is 0.72, and the proportion of trials where neither is visited is 0.39. As a direct result, we see more responses predicted at 100% than at 0% in Figure 4.2a, as it is more likely for state B to not be visited than state A. Only in 18% of the trials does the sampler actually reach both state A and B, meaning that in only 18% of the judgments a probability estimate is computed by comparing the nonzero frequencies of states A and B in the chain of samples. We will refer to these as ‘computed’ responses to contrast them from what we will refer to as ‘default’ responses at 0%, 50%, or 100% which occur when at least one state (A or B) was not visited by the sampler. The other 82% of the time the MS predicts such default responses at 0%, 50%, or 100%, which is why we observe large peaks in the dashed line at 0%, 50%, and 100% in Figure 4.2.

Let us now look at the effect of different chain lengths. As the predicted peaks at 0%, 50% and 100% by the MS are all due to certain network states not being visited by the sampling process, the number of samples drawn, i.e. the chain length, determines the size of these predicted peaks. Fewer samples drawn increases the probability that certain states are not visited and so larger peaks are predicted[14]. This effect of the chain length on predicted response distributions can be seen from the multiple red lines in Figure 4.2: Longer chain lengths, indicated by brighter red lines, have smaller peaks and provide a mean estimate that is closer to the normative answer. To assess the effect of chain length on the amount of default and computed responses we again simulated 10,000 runs with the MS, this time using chain lengths ranging from 2 to 48 (Table 4.1).

[14] The joint distribution of the causal variables also ties into this, since if there is a network state that has a small normative probability, it will be harder for the sampler to reach as well. This makes it that we observe the largest peaks in Figure 4.2a, where the query refers to a state where Y=1 and X_2=0, an unlikely state.

Predicted responses of the Mutation Sampler for $P(X_1 = 1|Y = 1,\ X_2 = 0)$

	Probability of response			
Chain length	Computed	0%	50%	100%
2	0.0	0.0177	0.927	0.0554
6	0.0605	0.0599	0.667	0.213
12	0.182	0.0936	0.392	0.332
24	0.418	0.0870	0.146	0.349
48	0.729	0.0309	0.0228	0.218

Table 4.1 *Predicted responses of the Mutation Sampler for the inference $P(X_1=1|Y=1,\ X_2=0)$. Probabilities are calculated by running the Mutation Sampler 10,000 times using the causal parameters of the common cause network in experiment 1A in Rottman & Hastie (2016). 'Computed' refers to responses that are computed by comparing the relative frequency of network states. We contrast these with 'default' responses at 0%, 50%, or 100%.*

As mentioned before, smaller chain lengths make it less likely that the required causal network states to compute an inference are visited by the sampler. With a chain length of two, the darkest red lines in Figure 4.2, the two states required to compute the inference are never both visited and the MS predicts responses to be exclusively at 0%, 50%, or 100%. From Table 4.1 we can indeed see that the $P(X_1 = 1|Y = 1,\ X_2 = 0)$ inference was never computed in 10,000 runs with a chain length of 2. For larger chain lengths these peaks are present as well albeit with a smaller magnitude. Even with a chain length of 48, which is at the high end of the range of chain lengths found previously (Davis & Rehder, 2020), the MS still predicts a substantial proportion of extreme responses at 0 and 100%. For instance, for the inference $P\left(X_1 = 1 \,\middle|\, Y = 1, X_2 = 0\right)$ the MS predicts responses at 100% for all chain lengths simulated (Figure 4.2a and Table 4.1). Even with a chain length of 48 the MS still predicts 22% of responses to be an extreme response of 100%. For each chain length above 2 the probability of an extreme response is at least 24% (Table 4.1). These predictions of the MS are not borne out, multiple studies have found participants to avoid the extremes of the scale in causal reasoning studies (e.g. Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016).

Figure 4.2 illustrates another aspect of the MS predictions that do not match the empirical data. We know that when the chain lengths are larger the predicted response peaks at 0%, 50%, and 100% decrease while the proportion of computed responses increases (Table 4.1). Regarding these computed responses, when the number of samples, i.e. the chain length, tends to infinity the predicted responses will tend towards the normative CBN point prediction. Hence, the mean prediction will get closer to the normative response with increasing chain lengths. One can see this happening from the red circles on the x-axis indicating the mean predicted response in Figure 4.2. This indicates that there is a tradeoff in the predicted distributions of the MS between the peaks at 0%, 50%, 100% and a peak of computed responses that gets closer to the normative response when chain lengths increase.

Based on this tradeoff the MS can predict mean conservatism by varying chain lengths. With very low chain lengths, the mean response is close to 50% as most responses will be default responses at 50%. At large chain lengths the mean response will approach the normative response. With more moderate chain lengths the mean response will lie in between 50% and the normative response (see circles on x-axis in Figure 4.2). This observation is consistent with the literature, as mean participant responses tend to be conservative and lie between 50% and the normative answer. However, it is not just the mean response that is between 50% and the normative answer. Typically, the bulk of responses tends to lie between 50% and the normative answer (Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). This however is inconsistent with the MS prediction, as the MS only predicts mean conservatism by trading off default responses (mainly at 50%) for computed responses (near the normative response). It is not able to predict the bulk of moderately[15] conservative responses in between 50% and the normative answer that have been found in experiments (Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). Relatedly, the model can predict some variability in responses that is similar to the empirical distributions (see predicted distributions with chain lengths 24 and 48 in Figure 4.2), however when it does so the mean predicted response tends to be off and there are still peaks that are not present in participant's responses. The mechanics of the MS lead to distributions that cannot mimic empirical responses in terms of certain distributional phenomena. To serve as a complete explanation of the cognitive process, the MS should be able to predict distributional behavioral phenomena in addition to mean phenomena.

This analysis of response distributions brought to light two important aspects of the data that the MS currently does not account for. The first is that it predicts extreme responses with a wide range of chain lengths. These responses are not observed in experiments, where people shy away from the extreme ends of the scale (Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). The second issue has to do with participant's conservatism: the bulk of responses is between 50% and the normative answer (Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). The mutation sampler can predict mean conservatism, but seemingly only by balancing the size of the peaks at 0%, 50%, and 100% with computed responses. It does not predict the bulk of responses to be in between the normative answer and 50%.

4.2.2 Forming judgments based on samples

To better understand how we could resolve the issues of the MS we identified above we consider the process of forming judgments based on samples. The most straightforward manner in which people can form probability judgments based on a set of samples is by calculating the relative frequency of an event occurring in the samples and taking this as an estimate of a probability of the occurrence of that event (Zhu et al, 2020). To illustrate this, imagine a scenario in which someone repeatedly throws tennis balls at beer bottles causing some to break. To estimate the probability that the bottle breaks with the next throw, we can compute the frequency of the bottle breaking in samples where a tennis ball is thrown:

$$Relative\ frequency = \frac{N_{breaks}}{N_{breaks} + N_{does\ not\ break}} = \frac{N_{breaks}}{N_{throws}}$$

[15] one can interpret 50% responses as 'extreme' conservatism

So, for instance, if bottles break 12 times in 20 throws, the relative frequency of the bottle breaking is $12/20 = 0.6$. Then, someone using the relative frequency approach would judge the probability of a bottle breaking as 0.6. This entails that a judgment is completely based on the incoming information and the judgment would approach the true probability when the number of samples tends to infinity. The MS uses this relative frequency method, computing judgments directly from the relative frequency of samples.

Problems with the relative frequency approach arise when we look at what happens when the number of samples is limited or small. According to the relative frequency method one could judge the probability of an event that is only witnessed once to be occurring 100% of the time. That is, if one observes a bottle to break after only one throw, one would judge the probability of a ball causing a bottle to break to be 100%. The reverse is also the case, the relative frequency method would lead one to judge anything that has not been directly observed yet to occur with a 0% probability. These extreme judgments are psychologically implausible. A more psychologically plausible model is to include prior information. The use of prior information can stop us from making extreme responses when we have little information to go by.

Besides preventing extreme judgments there is a more normative argument for the use of priors. The use of priors matches our decision-making in that it allows for gradual adjustments in the face of consistent evidence. When more and more tennis balls consistently break bottles, a judgment of 100% becomes more reasonable. This illustrates that people are sensitive to the amount of information obtained. When incorporating prior information, the amount of evidence presented (here the number of samples of throws, or ‘likelihood’ in Bayesian terms) does directly impact one’s judgment because we can weigh it relative to our prior information (e.g. our estimate gradually moves to 100% after seeing bottles break consistently). In contrast, the relative frequency approach would be insensitive to the amount of information learned from sampling.

Based on the previous theoretical arguments and the problems the MS has with predicting empirical response distributions, we propose the Bayesian Mutation Sampler (BMS) as an account of how people make causal judgments. The BMS is a process-model of causal reasoning combining mutation sampling (Davis & Rehder, 2020) with a generic Bayesian approach using priors to make probability judgments from samples (Zhu et al., 2020). We expect that the incorporation of priors will help in explaining the distributional behavioral phenomena discussed in previous sections.

In the next section, we will give a detailed overview of the BMS and subsequently will test whether it is an improvement over the MS, particularly in terms of the prediction of distributional properties, by fitting both models to experimental causal reasoning data.

4.3 THE BAYESIAN MUTATION SAMPLER

The BMS posits that when making causal probabilistic judgments people engage in sampling by way of mutation sampling (Davis & Rehder, 2020). This includes the principles of limited sampling and biased starting points, which bias judgments away from the normative CBN response. However, instead of using the relative frequency method to form judgments based on samples (as in the MS), the BMS incorporates prior information.

The type of prior information that people use for judgments and decision making varies. Many causal reasoning studies attempt to exclude the use of prior information regarding causal model parameters (e.g. Kolvoort, Fisher, et al., 2023; Rehder, 2014; Rottman & Hastie, 2016). However, even if researchers are successful in stopping participants from using prior information concerning causal parameters, it is likely that people still inherit priors relevant to the experimental task from similar everyday activities or in some way or another have expectations concerning the experimental task (see Hemmer et al., 2015; Marchant et al., 2021; Sanborn et al., 2021; Tauber et al., 2017; Welsh & Navarro, 2012). When specific task-related information is not present people can still use priors that reflect a lack of information.

The BMS posits that reasoners use a generic prior that encodes what they think to be likely answers to a causal probabilistic query before sampling. This prior gets updated based on the information in the samples. In Bayesian terms, the prior is updated using the information in the samples (the likelihood) to produce a posterior distribution. Subsequently probability judgments are based on this posterior distribution. Following Zhu and colleagues (2020), we take it that people respond using the expected value of this distribution (see also Jazayeri & Shadlen, 2010).

4.3.1 Incorporating the symmetric Beta prior

Following Zhu and colleagues (2020) we use symmetric Beta distributions as priors in the BMS, as they can reflect a lack of information in various ways and because they can be naturally incorporated into sample frequencies to form judgments.

Figure 4.3 plots symmetric Beta(β , β) distributions with values for β as the shape parameters. The Beta(1, 1) distribution is the uniform distribution, assigning equal probability mass to each probability p (from 0 to 1). For $\beta > 1$ the beta distributions assign more probability mass to the center of the scale, i.e. probabilities around .5. $\beta<1$ shows the opposite pattern, where more probability is assigned to the extreme ends of the scale. In this way using the symmetric Beta distributions allows the BMS to account for various levels of conservatism.

For all $\beta > 0$ the incorporation of the prior moves a response closer to 50% than when just using the relative frequency method. The only symmetric Beta distribution that would not introduce conservatism in this sense is the Beta(0, 0) distribution where all the probability mass is at the extremes of the range, at 0 and 1. Using the Beta(0, 0) distribution is equivalent to using the relative frequency method of forming judgments from samples. This entails that the BMS with β set to 0 is equivalent to the standard MS and so the BMS is a generalization of the MS.

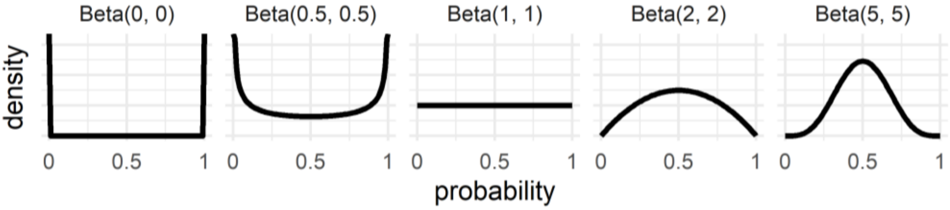


Figure 4.3 *Symmetric Beta(β, β) distributions, using β = 0, 0.5, 1, 2, and 5. Symmetric Beta distributions will be used as prior distributions in the BMS.*

By using the Beta(β, β) distribution as a prior, the expected value of the posterior distribution can be determined without computing the posterior distribution itself. We can compute the expected value directly by adding β as 'pseudo-observations' to EQ1 as in EQ2 (for the derivation of EQ2 we refer to Appendix A in Zhu et al., 2020)[16].

$$\hat{P}_{BMS}\left(X_1 = 1 \,\middle|\, Y = y, X_2 = x\right) = \frac{N(X_1 = 1, Y = y, X_2 = x) + \beta}{N(X_1 = 1, Y = x, X_2 = y) + N(X_1 = 0, Y = x, X_2 = x) + 2\beta} \qquad (EQ2)$$

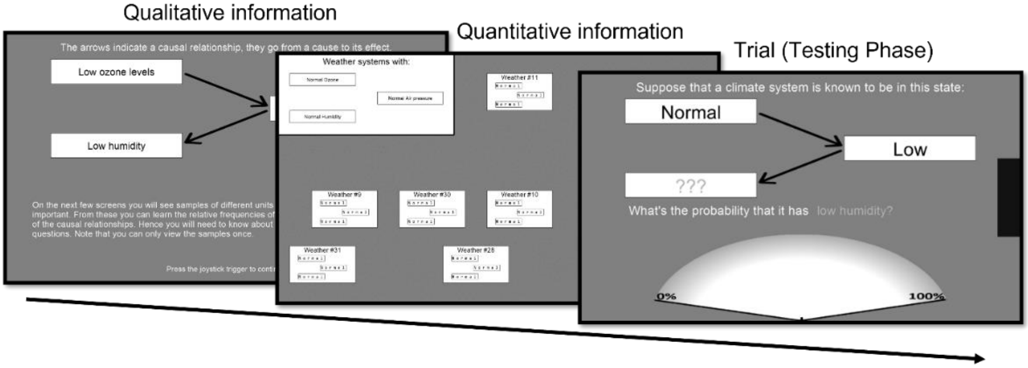
Here $\hat{P}_{BMS}$ refers to the estimate of the probability of an event predicted by the BMS, N stands for the number of samples (in a chain of generated samples), and X_i, Y refer to causal variables and x, y refer to their respective states. The β refers to the Beta(β , β) prior used, where both shape parameters of the Beta distribution are equal to β.

4.3.2 Testing the BMS

In order to validate whether the BMS provides a better explanation of response distributions than the MS, while still being able to predict mean responses as accurately as the MS, we fitted both models to data from a recent causal reasoning experiment (Kolvoort, Fisher, et al., 2023).

Here we provide a brief description of the experimental data, for a more detailed discussion we refer to the original paper (Kolvoort, Fisher, et al., 2023). The experiment consisted of three experimental domains, each comprising a learning phase and a testing phase. In the learning phase participants learned a specific causal structure, about which they were asked to make inferences in the testing phase (Figure 4.4).

[16] While conceptually different, our approach is computationally equivalent to one which would assign a prior probability to all the possible network states instead of to likely correct responses to queries. That is, if we would add β visits to all system states in EQ1 we would get EQ2.



The arrows indicate a causal relationship, they go from a cause to its effect.

On the next few screens you will see samples of different units
important. From these you can learn the relative frequencies of
of the causal relationships. Hence you will need to know about
questions. Note that you can only view the samples once.

Press the joystick trigger to conti

Weather systems with:

Suppose that a climate system is known to be in this state:

What's the probability that it has low humidity?

***Figure 4.4** Overview of experiment in* (Kolvoort, Fisher, et al., 2023) *with screenshots. First participants are taught about a particular causal system, receiving both qualitative (screen 1) and quantitative (screen 2) information on the causal variables, causal relationships, and causal strengths. Next, participants are asked to respond to (conditional) probability queries (screen 3).*

In the learning phase participants were provided with information about a causal system with three binary variables. They were given qualitative information concerning the variables and causal relations, as well as quantitative information using the experience sampling method with data sheets (Rehder & Waldmann, 2017) which involves participants viewing samples of data that manifest the statistical relations implied by the causal model. Each of the experimental domains used a different causal structure, which was either a chain, common cause, or common effect structure. The network parametrization of the structures was taken from Rottman and Hastie (2016, Experiment 1a), which was also used to fit the original MS to (Davis & Rehder, 2020).

In the testing phase participants responded to (conditional) probabilistic queries regarding the causal systems. Each of the 3 testing phases consisted of three blocks with different levels of time pressure implemented using response deadlines of 3, 9 and 20 seconds (of which the last one was intended to give participants ample time to respond). Each of these blocks consisted of 27 trials each consisting of a different inference. These inferences were of the form ‘Variable A has value x, variable B has value y. What is the probability that variable C has value z?’. Each of the three variables could have three states, one of the two binary values or unknown, leading to 3^3 = 27 different inferences. All participants completed 27 trials per domain and deadline condition, for a total of 27 x 3 x 3 = 243 trials. Participants responded on a scale from 0% to 100%.

Out of the 43 participants in the dataset, 17 did the study online. The only noteworthy difference between the online and offline study was the response modality; participants in the lab indicated a percentage by moving a joystick while online participants responded by moving their cursor using a mouse or trackpad.

We will fit the BMS and MS to each participant and condition (response deadline x causal structure) separately, this results in 43 x 3 x 3 = 387 sets of fitted parameters. In this way each set of parameters is fitted to 27 responses on 27 different inferences of a participant.

We did not identify a closed-form likelihood function for the BMS. Moreover, the parameters of the models consist of a combination of continuous (beta) and discrete (chain length) parameters. These considerations suggest a discontinuous or at least complex parameter landscape. A parameter recovery study (Appendix B) supported this suspicion, but revealed that using a parameter grid search (cf. Maaß et al., 2021; Mestdagh et al., 2019) resulted in correlations between true and estimated parameters consistently above .75 (see Method 2 Coarse grid in Appendix B), which is generally seen as good or excellent recovery (e.g. Anders et al., 2016; van Maanen et al., 2021; van Ravenzwaaij & Oberauer, 2009; White et al., 2015). Moreover, the parameter recovery study provided assurance regarding the identifiability of the BMS (Van Maanen & Miletić, 2021).

To fit the models using a grid search, we first simulate responses using the models with a range of realistic parameters (see below). These simulated responses were then saved in a grid. Each cell of this grid represents the predictions of the model under a particular set of parameters. To compute how closely the simulated responses match empirical responses we use the Probability Density Approximation method (PDA; Holmes, 2015; Turner & Sederberg, 2014) on each grid cell. PDA computes a 'synthetic' likelihood through kernel density estimation (Turner & Sederberg, 2014). The estimated parameters for a given condition and participant are from the cell with the highest likelihood given the data. We apply this method separately to each participant and 9 experimental conditions (3 levels of time pressure for each of the 3 causal structures) to obtain the optimal parameters for each.

To make sure that the grid contains the optimal parameters we set a wide parameter range. The chain lengths were varied between [2, 4, …, 68, 70], which includes the chain lengths found previously for reasoning tasks (Davis & Rehder, 2020) and the range of number of samples people are generally thought to generate (Vul et al., 2014). For the Beta prior parameter, we first included values for β from 0 to 1 with step size 0.1. Next, we included values for β > 1 based on the principle that the range of priors should be symmetric about the uniform prior. That is, priors with β > 1 would need to differ from the uniform distribution as much as priors with β < 1. To achieve this, we computed the total variation distance (Levin & Peres, 2017) between the uniform distribution and each prior in the grid with a β < 1. Then we identified the set of β > 1 that had the same total variation distance. This procedure resulted in the following betas: β ∈ [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.11, 1.26, 1.45, 1.73, 2.14, 2.83, 4.14, 7.35, and 21.54]. For β = 0 the principle of symmetry about the uniform prior would lead us to pick β = ∞. However, as using a Beta(∞,∞) prior would lead to responses only at 50%, we picked β = 100 instead.

In sum, we used a grid of 35 (values for the chain length parameter ranging from 2 to 70) by 21 (values for the β parameter ranging from 0 to 100) covering a wide range of plausible parameter values and simulated responses using (35 x 21 =) 735 different parameter combinations. While Davis & Rehder (2020) also estimated the causal parameters (base rates and causal strengths) of the causal structures that participants learned, we assume that participants learned the information they were presented accurately (we discuss this point further in the General Discussion). Hence with our setup the MS has only one free parameter (the chain length). The BMS has the β parameter for the symmetric Beta prior as a second free parameter.

Model code and helper functions to run simulations with the BMS and MS are publicly available at https://osf.io/xd9az/.

4.3.2.1 Overall fit

To quantify relative model performance of BMS to MS we computed BIC values for each set of fitted parameters (Schwarz, 1978). BIC, as compared to AIC, typically penalizes additional free parameters more strongly and so can be considered more conservative. We find that for 82.9% of the optimized models the BMS has a lower BIC value than the MS (mean Δ_{BIC} = -29.6). Next, we computed the average BIC weights per participant as approximations of posterior model probabilities (Neath & Cavanaugh, 2012; Schwarz, 1978; Wagenmakers & Farrell, 2004). We find that for each participant the BMS has a higher posterior probability than the MS (Figure 4.5).

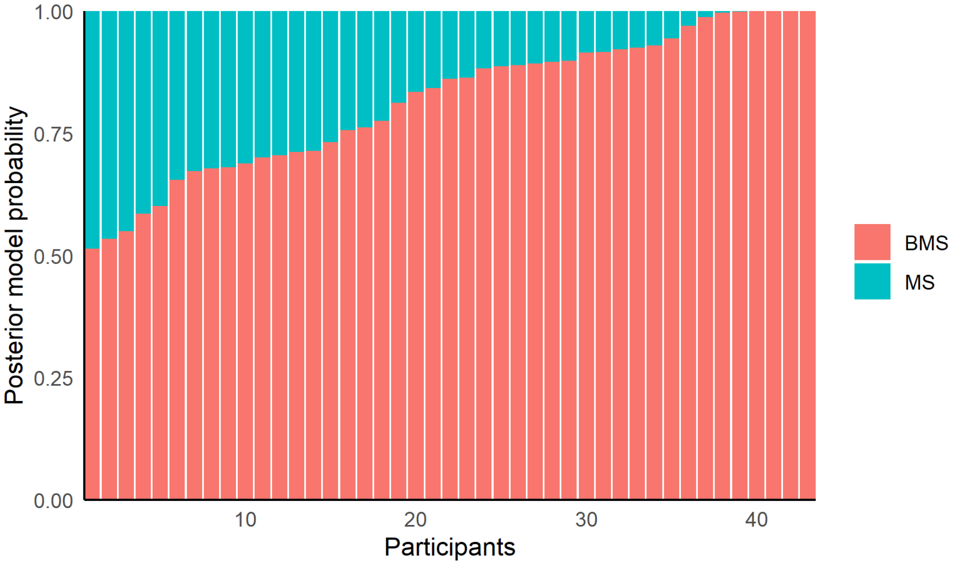


Figure 4.5 *Posterior model probabilities per participant for the BMS and MS. Posterior model probabilities are approximated using BIC weights.*

4.3.2.2 Mean predicted judgments

As discussed in the introduction, the MS accurately predicts mean responses on a variety of causal judgment tasks. To assess the mean predictions of the BMS we computed the expected value of all predictions from the BMS with the best fitting parameters of each participant. Specifically, per inference and per causal structure, we computed the average across participants of the predictions at each percentage point, resulting in an averaged predicted distribution, and then computed the expected value of this distribution. We find that the predictions closely follow the observed mean responses (Figure 4.6), indicating that the BMS is a good account of mean responses. The BMS outperforms the MS in this regard ($RMSE_{BMS}$ = 2.74; $RMSE_{MS}$ = 7.51).

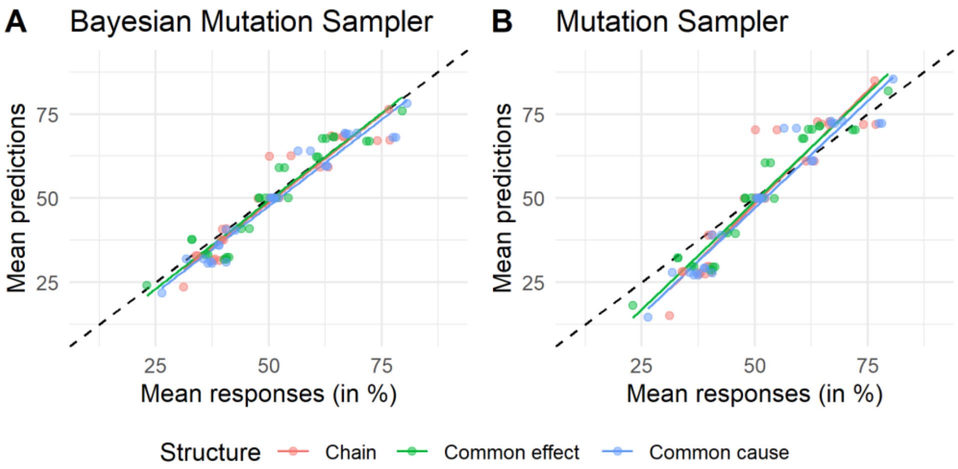


Figure 4.6 *Mean predictions of A. the BMS and B. the MS plotted against mean responses. Dots and lines are colored based on the causal structure in the experiment. Each dot represents one of the 27 inferences. Each line represents a linear fit to the predicted and empirical means. Black dashed diagonal lines indicate error free predictions.*

4.3.2.3 Variability of judgments

In addition to mean judgments, another important behavioral index is the variability of judgments (Kolvoort et al., 2021). However, getting a reasonable estimate of the variability in judgments is often challenging as it requires the repeated elicitation of comparable judgments (see Kolvoort et al., 2021). To obtain such repeated measurements and to present results concisely a common practice is to collapse over symmetry in the causal networks (e.g. Davis & Rehder, 2020; Kolvoort et al., 2021; Rehder, 2018; Rottman & Hastie, 2016). The joint distribution of the causal networks in the experiment used here were highly symmetric, allowing us to collapse over the terminal variables (e.g. $P(Y=1|X_1=1,X_2=0)=P(Y=1|X_1=0,X_2=1)$), over the presence or absence of variables by flipping responses to the upper half of the response scale (e.g. $P(Y=1|X_1=1,X_2=1)=1-P(Y=1|\,X_1=0,X_2=0)$), and over unknown variables (e.g. $P(X_1=1|\,Y=1)=P(Y=1|\,X_2=1)$). In addition, since we did not find significant differences in parameters, we will collapse over response deadline conditions. Finally, we will collapse over the chain and common cause network structures as these have an equivalent underlying normative distribution. We do not use the common effect structure for this analysis nor for the analysis of distributions below, since the small number of observations would lead to unreliable estimates of variability. Collapsing resulted in 7 groups of inferences presented in Table 4.2 (see Appendix C for an overview of all inferences in each group).

To index variability we use Gini's Mean Difference (GMD; David, 1968; Yitzhaki, 2003), defined as the average difference between any two observations. We use this non-parametric index as judgments on causal reasoning tasks tend to not be normally distributed (Davis & Rehder, 2020; Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rehder, 2018; Rottman & Hastie, 2016). To compute the GMD of model predictions, we first computed averaged predicted distributions for each participant and inference group (by averaging the predicted distributions over the deadline conditions, the chain and common cause networks, and the different inferences in each inference group). We then drew 10,000 samples from these aggregated distributions which we used to compute the GMD.

Inference groups

Group	Inference type	Conditioning information	Queried variable	Normative response	Obs. per participant	Example
1	Conflict	Two known variables with different values	1: terminal variable	75%	24	$P(X_1 = 1 \mid Y = 1, X_2 = 0)$
2			2: middle variable	50%	12	$P(Y = 1 \mid X_1 = 1, X_2 = 0)$
3	Ambiguous	Only one known variable	1: adjacent to known variable	75%	48	$P(X_1 = 1 \mid Y = 1)$
4			2: non-adjacent to known variable	62.5%	24	$P(X_1 = 1 \mid X_2 = 1)$
5	Consistent	Two known variables with the same values	1: terminal variable	75%	24	$P(X_1 = 1 \mid Y = 1, X_2 = 1)$
6			2: middle variable	90%	12	$P(Y = 1 \mid X_1 = 1\ X_2 = 1)$
7	Base rates	No known variables	-	50%	18	$P(X_2 = 1)$

Table 4.2 *Grouping of inferences for variability and distributional analysis based on symmetry in chain and common cause causal network structures. 'Queried variable' refers to the variable that participants are asked to judge the probability of. 'Terminal variable' refers to either X_1 or X_2 , and 'middle variable' to Y in Figure 4.1. The variable names in the example column refer to the variables as presented in Figure 4.1. See appendix C for a full list of inferences in each group.*

We find an empirical mean GMD of 13.8 indicating there is substantial variability in responses. Both models predict mean variability to be higher (GMD_{BMS} = 16.4, GMD $_{MS}$ = 19.2). The higher GMD for the MS is expected, because it predicts more extreme responses, increasing variability. Although the average variability of the BMS is higher than the observed variability, there are clear associations between the observed and predicted variability for each inference group (Figure 4.7 and Table 4.3). Table 4.3 presents the correlation coefficients of the predicted and empirical variability, for both BMS and MS.

Correlations predicted and empirical variability

	R	
	BMS	MS
Conflict trials 1	.72	-.19
Conflict trials 2	.58	.094
Ambiguous trials 1	.83	.40
Ambiguous trials 2	.74	.31
Consistent trials 1	.64	.37
Consistent trials 2	.68	.38
Base rates	.57	.38

Table 4.3 *Pearson correlations of predicted and empirical variability as indexed by GMD.*

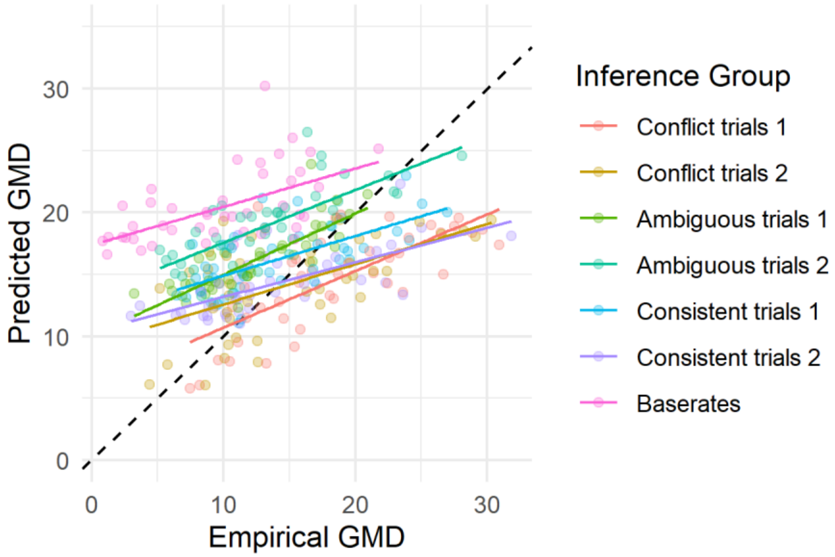


Figure 4.7 *Scatterplot of empirical variability and variability predicted by the BMS (indexed by Gini's Mean Difference, GMD). Responses and predictions are collapsed over the common cause and chain structures, the response deadlines, and into inference groups (see main text). Each dot represents one participant. Black diagonal indicates perfect predictions. Colored lines indicate mean linear trends per inference group.*

Within inference groups, the BMS predicts differences in variability between participants (Figure 4.7). However, the model does not perform well at predicting differences in variability between inference groups. For instance, it consistently predicts base rate judgments to be more variable than they are, and it predicts that judgments in the Conflict inferences 1 group are less variable than they are. That the BMS does not perform well in predicting between inference group variability might be due to that all different inferences are modelled with a single set of model parameters. While there seems no a priori reason that people use different priors for different inferences, it might be that the chain length differs based on the inference (see Gershman & Goodman, 2014; Hertwig & Pleskac, 2010; Vul et al., 2014; Zhu et al., 2020). When faced with a problem that is complex at first glance (e.g. an inference with conflicting conditional information), people could decide to sample for a longer duration. We return to this idea in the following sections. To get a better grasp of why some of the variability estimates are off we regard full response distributions next.

4.3.2.4 Distributions

To better understand the predicted distributions and how they match observed responses we present these distributions in Figure 4.8. Here the averaged best-fitting predictions of BMS and MS are presented together with histograms of participant responses.

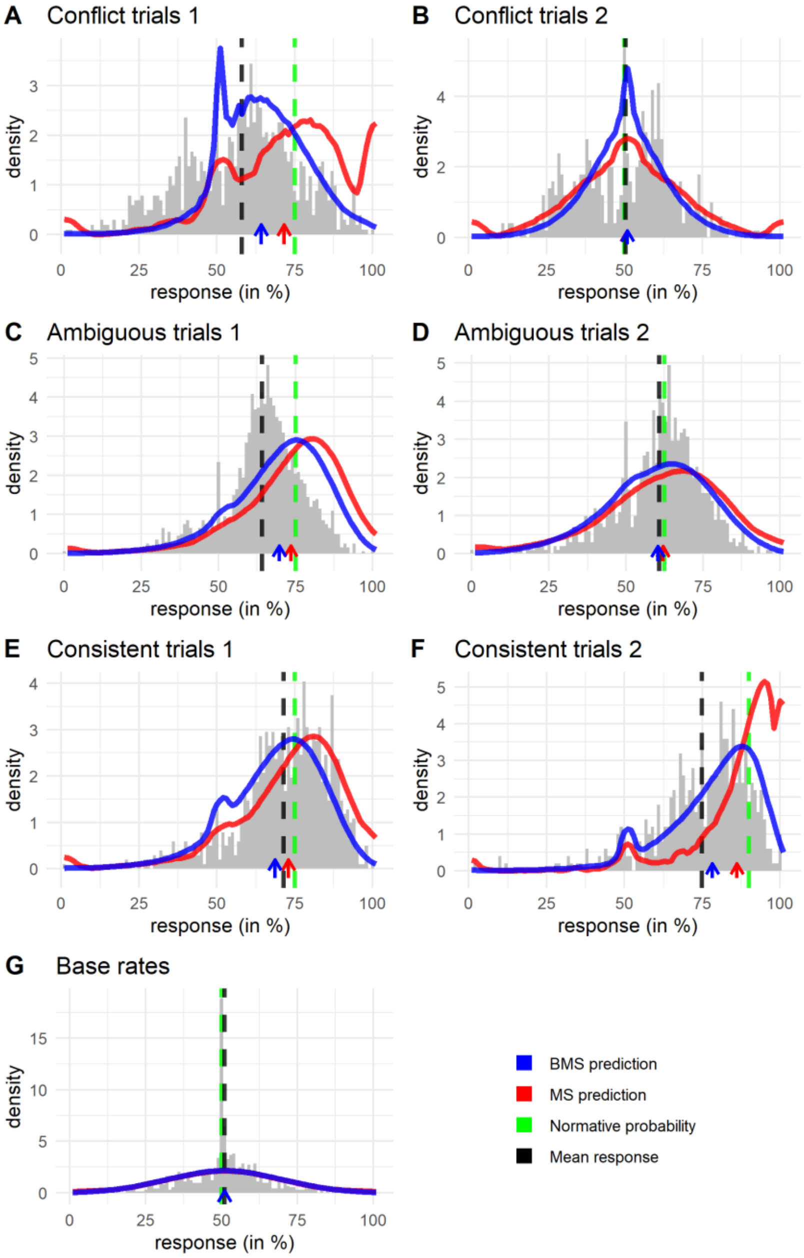


***Figure 4.8** Observed and predicted response distributions. Blue and red solid-colored lines indicate predictions based on model fits from the BMS and MS respectively, the arrows on the x-axis indicate the mean prediction. Grey histogram represents participant responses with the black dashed line indicating the mean. The green dashed line indicates the normative probability.*

First, let us discuss the distributional problems of the MS brought to light in the first part of this paper. The issue of extreme responses is visible relatively strongly in both the types of conflict trials (Figure 4.8 a and b). This can be expected, since conflict trials (where the conditional information is conflicting), require the sampler to visit mixed variable states. These states are harder to reach for the sampler, since the sampler is biased towards consistent, prototypical states. Hence the probability of a default extreme response is higher. Some extreme predicted responses by the MS are also visible for the ambiguous and consistent trials (Figure 4.8 c-f). The BMS does not produce any of these extreme responses. That there are fewer extreme responses for the MS than our analysis in the first part would indicate is because the chain lengths we found are higher than expected based on previous studies with the MS (Davis & Rehder, 2017; 2020). The second distributional issue we diagnosed of the MS, the lack of moderate conservatism, is also clearly visible in Figure 4.8 (panels a, c, e, and f). We see that the main mode of responses is at or more extreme than the normative probability. The BMS is more accurate in predicting where the bulk of responses are. Taken together, the BMS resolves the principal issues that the MS has in predicting where responses tend to be on the response scale.

The BMS is a clear improvement over the MS in predicting full response distributions and, according to our knowledge, the only process-level model of causal probabilistic reasoning that can satisfactorily account for response distributions. However, the predictions of the BMS are not perfect and certain limitations come to light when we regard the distributions for each inference type in more detail.

Figure 4.8a indicates that the BMS correctly predicts the mode of computed responses for conflict inferences where the middle variable is queried (Conflict trials 1). However, it wrongly predicts very few responses to be below 50%. Inspecting individual responses, we find that individual participants tended to respond on both sides of 50%. We observe this as well for the conflict trials where a terminal variable is queried (Conflict Trials 2, Figure 4.8b), hence it seems that the stimulus makes participants respond inconsistently due to the presence of conflicting cues. A substantial number of responses on this inference are possibly the result of random guessing which is not captured by the BMS. Such random responding might also explain why the frequency of responding at 50% is less than predicted. For the conflict trials where a terminal variable is queried (Conflict trials 2, Figure 4.8b) we see that the spike of responses at 50% is captured by the BMS prediction. What is not captured by the prediction is that participants tended not to respond close to 50%. One possible explanation for this observation is that participants rounded their responses to 50%. The reduced amount of responses near 50% occurs for both types of conflict trials, so it might be that the tendency to round to 50% is related to a participant's uncertainty in their estimates (cf. Kolvoort, Fisher, et al., 2023).

For the ambiguous trials, we observe that the prediction seems to be quite accurate for inference group 2 (Ambiguous trials 2, Figure 4.8d), but for group 1 (Ambiguous trials 1, Figure 4.8c) participants are more conservative than predicted. Such a pattern could be explained if participants sampled less, i.e. used a shorter chain of samples, to form a judgment in response to stimuli in group 1 than for group 2. When chain lengths are shorter, the influence of the prior is stronger and so responses would be closer to 50%. In our modeling we fixed the chain length over inference types. However, previous research has indicated that it is possible that people adaptively change their desired number of samples as the estimated costs and benefits of further sampling are dependent on the problem type (Gershman & Goodman, 2014; Hertwig & Pleskac, 2010; Vul et al., 2014; Zhu et al., 2020). Why might participants use different chain lengths for these inferences? Remember that for these ambiguous trials the state of one variable is given, while the state of the other non-queried variable is unknown. For the first group of ambiguous inferences the given variable is adjacent to the queried one, and so these stimuli may be considered as less ambiguous than group 2, where the given and queried variable are separated by an unknown variable. The observation that responses to stimuli in first group (Figure 4.8c) seem less variable than the second group (Figure 4.8d) is congruent with this idea. Since the stimulus in group 1 is less ambiguous, participants might view it as easier and so obtain fewer samples to form a judgment. A repeated measures ANOVA indicates that response times are indeed significantly shorter for group 1 than for group 2 (M_1 = 4.32s, M_2 = 4.62s, $F(39, 2891) = 14.73$, $p < .001$, BF = 32.5), corroborating this explanation. This effect is not captured in our predictions as we did not model a process by which participants might decide to use different chain lengths.

For the consistent trials (Figure 4.8 e and f) the BMS seems to capture the spread of responses rather accurately. For the base rate trials, however, the BMS severely underpredicts the spike of responses at 50% while capturing the rest of the variability quite accurately (Figure 4.8g). Compared to the other inference types participants have a strong tendency to respond at 50%. These base rate trials can be considered to involve the most uncertainty for participants compared to other inferences, since no conditioning information is provided. This uncertainty could lead to default responses or ‘guessing’ at 50% (Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). Upon viewing the stimulus participants might forego on sampling and instead respond at 50%. This would indicate a mixture of processes, where prior to the sampling process one might decide not to sample and instead respond in a default fashion. As we did not specify any mixture of processes the large spike at 50% is not captured by the BMS predictions here.

It is important to realize that we are putting up a very high bar when considering the ways in which the predicted distributions do not exactly match participant responses. The modeling of full response distributions is a complex endeavor as there are many processes and mechanisms that likely affect distributions of judgements which traditionally would be filtered out by taking the mean. In fact, many of the limitations of the BMS predictions just discussed point towards the need to specify additional processes to match the empirical distributions more accurately. We return to this point in the general discussion, before doing so we first regard the estimated parameter values.

4.3.2.5 Estimated parameters

A summary of fitted BMS parameter per response deadline condition is shown in Table 4.4. The overall median β parameter was 1.45, which is close to the uniform distribution (β = 1) that is often considered the prototypical uninformative prior. For most participants (79.1%) the mean β is larger than one, indicating they expected answers to be closer to 50%, validating our choice to include values for β > 1 (cf. Zhu et al., 2020). Higher values for β lead participants to be more conservative in their responses. This could explain (a part of) the substantial conservatism observed on this (Kolvoort, Fisher, et al., 2023) and other causal reasoning tasks. While no participant was fitted best by the upper bound of the grid (β = 100), we find that in 5.94% of cases the best fitting β is zero, the lower bound of the grid. This indicates that in only a small subset of cases the relative frequency method of generating judgments (as proposed by the original MS) was used. As could be expected the prior used was not affected by response deadlines (F(2, 336) = 1.04, p = .433, BF_{H1} = 0.036).

Summary fitted BMS parameters

	β parameter			Chain length		
Deadline	Median (*SD*)	Minimum	Maximum	Median (*SD*)	Minimum	Maximum
6s	1.73 (2.58)	0	21.5	40 (19.0)	8	70
9s	1.11 (2.77)	0	21.5	46 (19.5)	4	70
20s	1.45 (3.14)	0	21.5	50 (20.4)	6	70
Overall	1.45 (2.83)	0	21.5	44.9 (19.7)	4	70

Table 4.4 *Summary of fitted BMS parameters based on fitting to each participant, structure, and deadline condition separately, resulting in 387 sets of parameters. The β parameter refers to the fitted Beta(β, β) priors.*

The average chain lengths we find fall within a range expected based on the literature. Zhu et al. (2020) for instance found best fitting chain length for certain participants to be well over 200 in simple probability judgment tasks. While Davis & Rehder (2020) found a maximum mean chain length of 28 for causal inference studies, the average best fitting chain length for some causal intervention or categorization studies was above 60. There seems to be a trend of increasing chain lengths for longer deadlines (Table 4.4). This would be consistent with the hypothesis that participants sample longer when they have more time to respond. However, a repeated measures ANOVA indicates that there is no statistical support for such an effect (F(2, 336) = 3.74, p = .121, BF_{H1} = 0.25). 14.0% of fitted chain lengths reached the maximum value of 70. This reflects that larger chain lengths are more difficult to estimate, since the differences in the predictions of the BMS become increasingly smaller as chain lengths increase (Appendix B). While there is some uncertainty regarding the exact values of the higher chain lengths, the median chain lengths we find are noticeably higher than found by Davis & Rehder (2020) for experiments involving causal inference, as they found the best fitting chain lengths to range from 4 to 28 in these types of tasks[17].

4.3.2.6 BMS and behavioral measures

Lastly, we studied the relationship between the fitted parameters and other behavioral measures to validate the BMS. As the BMS is a process model, it should relate to behavioral measures not specified in the model itself. We looked at three important behavioral measures: response times, accuracy, and conservatism. Response times here are especially of interest as they are not part of the data used to fit the BMS. Accuracy was defined as the absolute distance of responses from the normative answer and conservatism was defined as the absolute distance of responses from 50%. For ease of interpreting the statistics, we multiplied both accuracy and

[17] The uncertainty regarding the exact values of higher chain lengths does not affect the conclusion that we found higher chain lengths, since if participants indeed used only few samples (smaller chain lengths) to make judgments, our parameter recovery study indicates we would have recovered those parameter values accurately (Appendix A).

conservatism values with minus one, as by doing so higher values indicate higher levels of accuracy and conservatism respectively. We computed the mean fitted parameters and the means of these behavioral measures over all conditions per participant and tested their relationships using Spearman correlations (Table 4.5).

Correlations fitted parameters and behavioral measures

	Response time	Accuracy	Conservatism
Chain length	.344*	.893***	-.323*
	(3.29)	(>100)	(2.48)
β	-.0974	.00740	.749***
	(0.405)	(0.341)	(>100)

Table 4.5 *Spearman correlations of mean fitted parameters and behavioral measures per participant. See main text for definitions of Accuracy and Conservatism. Numbers in brackets refer to Bayes factors (*BF_{H1}*, i.e. for the existence of a correlation), computed using the BayesFactor package in R using default settings (Morey & Rouder, 2014). Asterisks indicate p-values: *$*p < .05$*, *$**p < .01$*, *$***p < .001$*.*

We find that chain length is positively correlated with response times and strongly with accuracy, while being negatively correlated with conservatism. These correlations all reflect a higher task performance of individuals that sample for a longer duration: Firstly, as the chain length is a direct reflection of the sampling duration, a correlation with response time is expected. Secondly, longer chain lengths indicate more computed responses that get to the normative answer as the effect of the starting point bias decreases, resulting in higher accuracy. Finally, longer chain lengths also imply less influence of the prior, which is reflected in responses being less conservative. This is also reflected by the strong positive correlation between the β parameter and conservatism. Higher values of β imply the use of a prior that has more probability mass near 50% which results in more conservative responses. In sum, these correlations support the BMS model specification by showing expected relations between the parameter estimates and behavior.

4.4 GENERAL DISCUSSION

The aim of this paper was to understand distributions of probabilistic causal judgments. In the first part we diagnosed problems with the distributions of responses that a process model of causal reasoning, the Mutation Sampler (MS), predicts. The two main problems we identified were that the MS predicts a non-trivial number of extreme responses at 0 or 100 %, and that it predicts the bulk of computed responses to be centered near the normative probability. Contrary to these predictions, data indicate that people actually refrain from using the extreme ends of the response scale and that the bulk of their responses tends to lie in between the normative answer and 50% (Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rehder, 2018; Rottman & Hastie, 2016). We traced these issues back to the process by which the MS forms judgments based on samples and proposed to extend the MS by the incorporation of prior information into judgments. In the second part of the paper, we formalized the idea of incorporating prior information into judgments and presented the Bayesian Mutation Sampler (BMS). The BMS combines the sampling process of the MS with the use of prior information. We fitted the BMS and MS to data from a recent causal reasoning experiment to illustrate that the BMS provides a better account of the data, and found that the BMS resolves the distributional problems associated with the MS. Although the BMS predictions are not perfect, the model is able to account for a lot of the variability we observe in causal judgments. To our knowledge the BMS is the only computational (process-level) model of causal reasoning that is able to capture response distributions to this degree. This is not an easy feat, especially considering that the model uses only two free parameters to predict full response distributions for multiple inference types. These findings provide evidence for the notion that the variability observed in causal judgments is due to a sampling mechanism. This is in line with findings implicating sampling being the source of variability in children's responses on causal learning tasks (Bonawitz, Denison, Griffiths, et al., 2014)

Our formulation of the BMS entails some important theoretical commitments. We proposed that participants have prior beliefs regarding likely (conditional) probabilities and that they incorporate these beliefs into their judgments in Bayesian fashion. Amongst other reasons, we motivated the use of priors by the observation that participants do not provide extreme judgments. That is, we interpret the avoidance of the extremes of the response scale as a rational adjustment to small sample sizes via the incorporation of prior information. However, there are other plausible explanations for the observation that people avoid making extreme judgments. One option would be that people avoid the extremes of the response scale due to a response bias; e.g. participants could be reluctant to express the confidence that a judgment of 0% or 100% might imply (e.g. DuCharme, 1970; Phillips et al., 1966). Another option could be that people use a particular mapping of objective probabilities to subjective probabilities, such as a probability-weighting function used in Prospect Theory (Tversky & Kahneman, 1992). The BMS posits that people avoid the extremes of a probabilistic response scale due to their beliefs (encoded by priors) instead of this being the result of a particular mapping from beliefs to a response scale. While we are not necessarily committed to the idea that participant beliefs map straightforwardly onto a probabilistic response scale, we did not implement a mapping function.

A different approach was used in the original MS paper, where the authors used a scaling factor to map the probabilities computed by the MS to the response scale (Davis & Rehder, 2020). A free 'scaling' parameter *s* was used such that a predicted response = s*p, where p is the probability generated by the MS. The MS combined with such a mapping function can produce responses that fall outside the response scale and so we did not consider it to be a proper account of the variability in causal judgments (Appendix D provides an illustration of this effect). Including such a scaling parameter improves the fit of the MS, but it is still outperformed by the BMS (Appendix D). Moreover, as the use of a (Beta) prior is theoretically motivated the BMS is clearly the favored model.

We can consider the alternative explanations for conservatism mentioned earlier as other theoretically motivated mechanisms that could map probabilities produced by the MS to a response scale. In other words, a response bias related to a reluctance to express confidence or the use of a probability-weighting function as in Prospect Theory could be assumed instead of a Beta prior. These explanations could predict that responses are 'pushed' towards the middle of the response scale in a similar way as the incorporation of a symmetric prior does. Due to this we cannot use the distribution of responses in the current experiment to distinguish between these explanations. One possible approach to empirically verify the prior mechanism proposed by the BMS would be to conduct an experiment in which one manipulates participant beliefs about what the likely answers to a causal probabilistic query are. To manipulate prior beliefs participants need to be presented with data. This data could be of various forms, such as training data that imply extreme probabilities, fabricated responses from other participants, or participants could be provided feedback on their own responses. If the BMS is correct we would expect participants to update their prior in light of this new data which in turn would systematically affect their judgments. We suggest future research to run such experiments to confirm the use of priors. As it currently stands, though, we maintain that the use of priors is the most plausible explanation of the observed behavior as there are compelling arguments in its favor that go beyond response mapping. There are the usual normative arguments in favor of a Bayesian approach and people have been shown to reason in a Bayesian manner (that is, using priors) in many other domains (see Oaksford & Chater, 2020; Parpart et al., 2018; Vul et al., 2014; Zhu et al., 2020).

In all, our work contributes to a recent movement in the field of causal reasoning arguing that the variability observed on tasks reflects information of interest (Kolvoort et al., 2021; O'Neill et al., 2022). Understanding the variability in responses and modeling full response distributions can be a challenging task but it comes with important benefits (see O'Neill et al., 2022). To promote future research in this direction the remainder of this paper discusses potential pitfalls and gains of such an approach and suggest promising directions of research into causal judgments.

4.4.1 The importance of distributions

Important benefits of shifting the explanatory focus from mean responses to response distributions include the potential to ask more questions and providing safeguards against drawing erroneous conclusions.

4.4.1.1 Asking more questions

Recent examples of using response distributions to better our understanding of causal reasoning include initial preliminary investigations by Rehder (2018) and Rottman & Hastie (2016), and the more recent studies by Kolvoort et al. (2021) and O’Neill et al. (2022). Using a novel experimental design Kolvoort et al. (2021) elicited repeated causal judgments which allowed them to establish the presence of substantial within-participant variability that differs per inference type. O’Neill et al. (2022) analyzed response distributions of existing and new vignette-based experiments which led them to conclude that causal judgments are often graded and multimodal. This result allowed them to assess theories of causal reasoning and suggest improvements based on a graded concept of causation. Similarly in our work, both predicted and empirical distributional phenomena informed the development of the BMS. The observation that the MS predicts extreme responses and many responses near the normative probability was crucial in determining how the MS could be improved. A shift towards analyzing distributions will allow researchers to target more behavioral phenomena and subsequently develop more comprehensive theories.

4.4.1.2 Stop erroneous inferences

In addition to leading us to more questions, using distributions instead of means as our explanatory target can help us to not draw erroneous conclusions. It has become clear from the current work and recent investigations that the variability of causal judgments does not just reflect noise and that these response distributions are often multimodal and non-normal (Kolvoort et al., 2021; O’Neill et al., 2022; Rottman & Hastie, 2016). This entails that what we infer from statistical and cognitive models which characterize only mean responses can be severely misleading. This can be illustrated by looking at what would happen if we had merely modeled mean responses using the BMS. Figure 4.9 shows three predicted distributions for a single inference of the BMS using different sets of parameters. While the chain lengths range from 18 to 70 and the β parameter from 0.1 to 2.1, the mean response for these distributions is the same. Hence the model would not be identified if we were to only regard mean responses. Moreover, these distributions, though having the same mean, imply a very different type of responding (cf. Anders et al., 2016). For instance, while it is common to assume that most responses are near the expected value of a response distribution, Figure 4.9 shows that the BMS can produce varying densities of responses near the expected value while keeping the expected value itself fixed. This is an issue related to model identifiability (cf. Van Maanen & Miletić, 2021) and it is quite common outside of the domain of causal reasoning (for instance with models predicting response time distributions, e.g. Anders et al., 2016). For the BMS, increasing the chain length moves the expected value towards the normative response, while increasing β moves the expected value of the predicted distribution towards 50%. Hence for any percentage point Z that is between the normative probability and 50%, there are an infinite number model parameter combinations that would produce distributions with the expected value at Z. Due to this solely focusing on the mean response can lead to drawing multiple erroneous conclusions. One wrong conclusion could be that most responses lie near the mean predicted response, which is often not the case for causal judgments (e.g. O’Neill et al., 2022). Relatedly, the amount of disagreement between participants could be overestimated if one were to erroneously conclude that the grey prediction in Figure 4.9 accurately captures group-level responses (compared to other predictions in Figure 4.9).

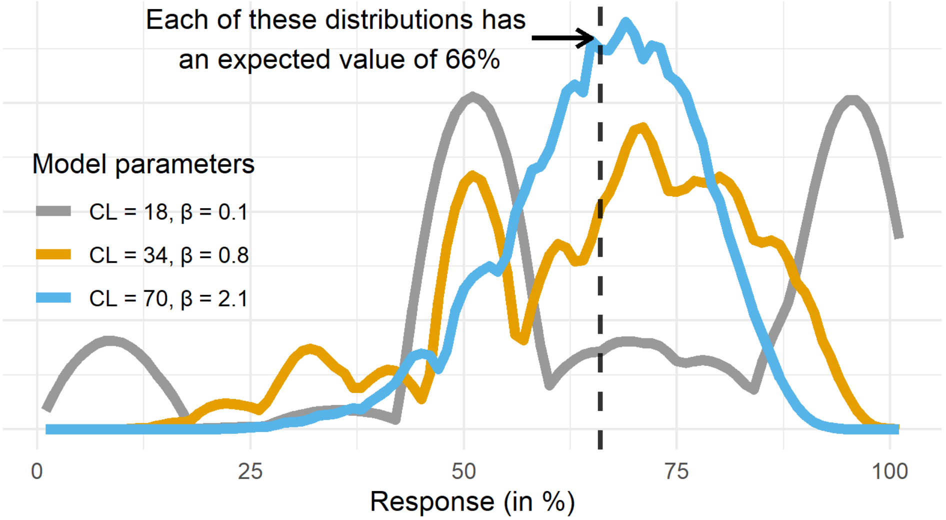


Figure 4.9 *Illustration of BMS predicted response distributions of the inference* $P(X_1 = 1|Y = 1, X_2 = 0)$ *for 3 different sets of parameters using a chain causal structure with the same parametrization as used in the experiments studied in this manuscript. 10,000 simulations were run to compute each of the three distributions. Each of the distributions was smoothed using a kernel density estimate as specified by the PDA method. CL refers to the chain length parameter, β refers to the parameter for the Beta(β, β) prior. The vertical dashed line indicates the expected value of all three distributions.*

To make sure researchers do not draw erroneous conclusions based on models that overemphasize the importance of the central tendency of responses, O’Neill et al. (2022) recommend to plot histograms of response data regularly and to assess whether the underlying distributional assumptions of their (statistical) models are met. This latter point is important, as the standard linear models often assume equal variances over participants and conditions, while we now know that this assumption is often violated in causal judgments (e.g. Kolvoort et al., 2021; O’Neill et al., 2022).

4.4.1.3 Using generative models to target distributions

Another recommendation O’Neill et al. (2022) give is for researchers to move towards modeling response distributions using a generative approach. This computational approach is becoming more widespread in psychological and brain sciences (see for recent overviews Ahn & Busemeyer, 2016; Forstmann et al., 2016; Guest & Martin, 2021; Jarecki et al., 2020; M. D. Lee et al., 2019; Ratcliff et al., 2016; Turner et al., 2017; Wilson & Collins, 2019). Generative modeling involves constructing computational models that embody theoretical assumptions about how behavior is generated (Haines et al., 2022). This involves characterizing the psychological process that turns inputs (stimuli) into outputs (behavior or judgments) in mathematical terms.

The BMS (as well as MS) is a generative model of the process by which people generate causal probabilistic judgments by way of sampling from a causal network (Davis & Rehder, 2020).

Generative modeling naturally leads researchers to focus on psychologically interpretable parameters (such as the number of samples or type of prior information) instead of on estimating descriptive effects. Such descriptive effects are often defined using differences in mean response between conditions and tested using standard parametric statistical tools (such as *t*-tests, regressions, ANOVAs, etc.). However, we know that means often don't capture response distributions in a meaningful way and that this violates the assumptions of the standard parametric tests (see Haines et al., 2022). These descriptive effects could also be modeled using 'descriptive' models that predict only mean responses (e.g. Rehder, 2018; Rottman & Hastie, 2016), but this could lead one to draw erroneous conclusions as the mean response could misguide researchers (see Figure 4.9).

Instead, generative modeling helps researchers to account for more than just averaged behavior. When mathematically specifying the data-generating process a researcher has to incorporate assumptions about the psychological processes that generate empirical data. These assumptions should allow the model to generate predictions that mimic not just empirical means, but empirical distributions.

While generative models can be used to characterize group-level behavior, often they are implemented at the level of individual psychological processes. This allows for the fruitful study of individual differences, which could help in assessing competing models of causal reasoning (a good example can be found in Gerstenberg et al., 2021). We did not focus on explaining such individual differences here, but the BMS can in principle explain inter-individual variation by appealing to differences in how long people sample and the prior information they use. Since it is a longstanding question in the field of causal judgments to what extent the observed variability is due to within- or between-participant variability (Davis & Rehder, 2020; Kolvoort et al., 2021; Rehder, 2018; Rottman & Hastie, 2016), this seems to be an important direction for future research.

Related to individual differences, note that Davis & Rehder (2020) did not fix the causal parametrization of the causal network in the MS, instead estimating the causal parameters (i.e. base rates of the causal variables and causal strengths) to account for participants not learning an accurate representation of the causal network. Consequently, there is no strict one-to-one correspondence between their results and ours. The justification for estimating causal parameters was that we cannot assume that participants' representations actually conform to what they are taught during the experiment about the causal networks. There is some force behind this argument as it is unlikely that participants learn the qualitative and quantitative aspects of the causal systems exactly as how they are presented to them. However, the current study was not aimed at understanding such individual differences. As mentioned, the current implementation of the BMS could explain individual differences by appealing to differences in how much people sample and the prior information they use. It might be that other individual factors need to be estimated to capture individual differences, such as individually learned causal parameters or, for instance, individual subjective probability weighting functions (Tversky & Kahneman, 1992) or differences in the order of processing of variables (Mistry et al., 2018). Which factors explain individual differences in probabilistic causal judgments remains an open question for now.

4.4.2 Overcoming challenges in studying distributions: extending the BMS

The study of full response distributions, possibly using generative process models, is a promising direction and will help advance our understanding of the cognitive processes responsible for causal judgments. This type of work does, however, come with its own set of challenges.

4.4.2.1 Modeling the multitude of processes resulting in a response

One possible direction to improve the BMS would be to adapt its core features. It could be that a different sampling mechanism (e.g. a Gibbs sampler instead of the MH algorithm; cf. Davis & Rehder, 2020) or differently shaped prior could help explain more of the variability. However, our results seem to indicate that the limitations of the BMS predictions are due to additional processes at play. That is, to capture empirical response distributions more accurately we need to model these processes.

The existence of additional processes affecting responses is a big challenge in trying to account for response distributions. There are a multitude processes that affect how judgments are made. Examples of such processes are the rounding of estimates (e.g. Budescu et al., 1988; Costello & Watts, 2014; Erev et al., 1994; Kleinjans & van Soest, 2014), guessing (Kolvoort, Fisher, et al., 2023; Kong et al., 2007; Schnipke & Scrams, 1997), a general mixture of multiple problem-solving strategies (Archambeau et al., 2022; Evans, 2008; Van Maanen et al., 2014, 2016) and 'dynamic effects' such as fatigue, boredom, and learning (e.g. Gunawan et al., 2021). All these processes affect participant's judgments on experimental tasks and partly determine the resulting response distribution.

The existence of all these additional processes is an important rationale in studying mean judgments: averaging across many trials filters out supposed noise. However, when targeting response distributions with generative models this solution is not available. Instead, we need to tackle the inherent complexity and develop theories and models explaining the target distributional phenomena.

While there are many processes that can be modelled, we focus here on two that our results indicated might play a large role in determining the observed distributions. The first is using a mixture of strategies. One likely strategy that is consistent with our observations is a guessing strategy. We observed varying peaks of responses at 50% which have been attributed to guessing before (Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). A second alternative strategy is that of adapting chain lengths based on the inference type one is confronted with. Such a mechanism has been suggested before (e.g. Zhu et al., 2020) and could, for instance, explain the difference in fit between the two types of ambiguous inferences (see Results section). While there can be many other processes affecting responses, we hypothesize that these two are likely to account for a substantial amount of the unexplained variability.

4.4.2.2 Modeling mixtures

People can employ a variety of strategies to perform a particular task and so it is often assumed that observed behavior is the result of a mixture of such strategies (Campbell & Xue, 2001; Coltheart et al., 2001; Couto et al., 2020; Donkin et al., 2013; Dunlosky & Hertzog, 2001; Dutilh et al., 2011; Evans, 2008; D Kahneman, 2011; Smith et al., 1998; Van Maanen et al., 2014; Van Maanen & Van Rijn, 2010; Van Rijn et al., 2003). For example, the dual-process framework of decision-making is built on the idea we can solve problems in either a more intuitive/heuristic manner or in a more deliberative manner (Evans, 2008; D Kahneman, 2011; Stanovich, 1999). Intuitive reasoning is characterized by being automatic, fast and non-conscious, which is contrasted with deliberative reasoning that is more rule-based, slow, analytic and controlled. We can see this in two ways related to the BMS. One way is to consider the BMS to implement both intuitive and deliberative reasoning. If sampling chains are short responses are most affected by the biased starting points (prototypes), which can be seen as intuitive. When sample chains are long, the response is based more on the learned statistical relationships and we can consider this deliberative. Another way to look at this is to consider all sampling (e.g. as implemented by the BMS) as a deliberative way of reasoning. In this case, intuitive reasoning would be implemented by a wholly different mechanism. For instance, a more heuristic manner by which participants could respond on causal reasoning tasks would be by using a simple 'tallying' rule (Gigerenzer & Gaissmaier, 2011; Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016), which involves counting positive cues (present variables) and subtracting negative cues (absent variables) to form a judgment. Even simpler strategies would be to just guess 50% or respond randomly, which is believed to happen on a subset of trials (Kolvoort, Fisher, et al., 2023; Kong et al., 2007; Schnipke & Scrams, 1997).

There is evidence that participants use such simple strategies in causal judgment tasks. One salient feature of response distributions in the probabilistic causal judgment literature are spikes of responses at 50% and it has been found that the size of these spikes differs per inference type (Kolvoort et al., 2021; Rottman & Hastie, 2016). Our results are in line with these findings. We found spikes of responses at 50% whose size depended on the inference type (see Figure 4.8). The large spike at 50% for base rate judgments stands out the most as it is substantially larger than the BMS would predict (Figure 4.8g).

We believe that this large spike at 50% for base rate judgments may reflect guessing or a 'default' response at 50%. In fact, spikes at 50% have previously been associated with uncertainty in the response (Kolvoort, Fisher, et al., 2023; Rottman & Hastie, 2016). When participants are uncertain or do not know how to respond they could employ their default response strategy and respond at (or around) 50%. This reasoning might explain why the spike at 50% is highest for base rate trials. For base rate trials the stimulus is most uncertain, as no information regarding the other variables in the causal structure is given. Hence based on this uncertainty explanation we would expect to see many such 50% responses for base rate queries.

While our modeling approach assumed that all responses were generated by way of mutation sampling, it is likely that participants varied in how they responded and that in a subset of trials they just guessed 50%. Additionally, for Conflict trials 1, in which the queried variable is adjacent to two opposing cues, we found that individual participants responded on both sides of 50%. This type of responding is consistent with a strategy of random responding.

If indeed the observed response distributions on causal judgment tasks are the result of a mixture of strategies, then it appears that the mixture proportions are dependent on the inference type. One piece of evidence for this is that we found substantially larger spikes for base rate inferences. More evidence come from previous studies, both Rottman and Hastie (2016) and Kolvoort et al. (2023) concluded that the frequency of 50% responses depended on the uncertainty that reasoners might have. That is, they found spikes to be smallest for consistent inferences, larger for ambiguous inferences, and largest for inconsistent inferences. We also observe this pattern in our data (Figure 4.8). These findings seem to point toward a mechanism in which people resolve their uncertainty by responding using the middle of the response scale (Kolvoort, Fisher, et al., 2023). It suggests that people adapt their response strategy based on an uncertainty-related feature of the stimulus. People use the 'guessing strategy' more often when there is a lot of uncertainty (e.g. a conflict or base rate trial) versus when there is not (e.g. a consistent trial).

4.4.2.3 Adaptive chain lengths

Another potential source of variability that comes to light once we start modeling the full response distributions is that possibility of variable chain lengths. That is, it may be that the number of samples one considers for making a judgment, could differ per inference type. We found that for the ambiguous trials, where the state of one variable was unknown, participants were more conservative and responded more quickly when the known variable was adjacent to the queried one. This is consistent with the effects of having a shorter chain length. In the Results section we proposed that this might be due to these inferences being less ambiguous leading to people thinking they can be relatively accurate without needing to generate many samples.

While we could estimate the chain length separately for each inference type, a more principled approach would be to determine why and how chain lengths differ and to incorporate this into the BMS. Future research could focus on investigating this relation between stimulus and chain length. Zhu et al. (2020) suggest that how much someone samples could be dependent on problem complexity. This would be consistent with our ambiguity-based explanation. Related to this idea, other researchers have proposed an adaptive scheme in which the costs and benefits of samples are weighed to determine how many samples to generate (Gershman & Goodman, 2014; Hertwig & Pleskac, 2010; Vul et al., 2014; Zhu et al., 2020). These seem like fruitful ways to extend the BMS.

4.5 CONCLUSION

We studied the predictions of the MS, as it currently is the most promising process-level model of causal reasoning, and found it has some shortcomings in explaining full response distributions. The original MS model was developed to implement four psychological principles that apply to causal inference: we think about concrete cases, we make small adjustments to these cases, we have a bias for prototypes, and we can only draw a limited number of samples. By developing the

BMS and showing its improved performance we have argued for an additional principle to be added to this list: people make use of prior information. By adding this principle, and implementing it with the BMS, we showed it is possible to account for more distributional phenomena in causal judgments. We hope this work spurs other researchers to focus efforts on analyzing more than just mean responses as we believe this will improve our understanding of underlying cognitive mechanisms greatly.

5 MODELS OF VARIABILITY IN CAUSAL JUDGMENTS

Abstract

Most theories of causal reasoning aim to explain the central tendency of causal judgments. However, experimental studies show that causal judgments are rather variable and that this variability is informative. The current study investigates the extent to which multiple candidate theories of causal reasoning explain such variability in causal judgments. To this end, we implement computational cognitive models of these theories and fit those to data from a previously published experiment that includes repeated probabilistic causal judgments. We find that the Bayesian Mutation Sampler provides the best account of the data. This suggests that the stochastic sampling mechanism posited by the Bayesian Mutation Sampler is an important source of variability in causal judgments. Additionally, our findings suggest that incorporating 'non-reasoning' processes, such as rounding and guessing, into models of causal reasoning can improve their ability to account for the observed response distributions. Overall, the study highlights the potential of computational modeling to shed light on the underlying mechanisms of human causal reasoning and identifies promising directions for future research in this domain.

This chapter has been adapted from: Kolvoort, I.R., Davis, Z.J., Rehder, B., & van Maanen, L. (Manuscript in preparation). Models of Variability in Causal Judgments.

5.1 INTRODUCTION

One important way in which we understand the world is through the lens of causation. Our knowledge about causality in our environment has been found to affect a myriad of decisions and judgements (see Danks, 2014; Sloman, 2005; Sloman & Lagnado, 2015; Waldmann, 2017b). Over the last decades a renewed interest in causal cognition has led to a wealth of studies investigating different facets of causal cognition. One of the main tools used to understand human causal cognition is a theoretical framework known as causal Bayesian networks[18] (CBNs; Pearl, 2009; Spirtes et al., 2000). As a normative theory CBNs have provided a decent approximation of human behavior and it has provided researchers with a benchmark with which to compare human behavior. While an important tool, CBNs in themselves do not provide us with knowledge about *how* causal judgments are being made. Instead, they provide a computational account that allows us to formally describe what causal judgments people make and it allows to distinguish between those. However, as cognitive scientists and psychologists it is of great interest to us *how* causal judgments are made, i.e. to understand what cognitive processes lead to these sophisticated judgements.

Before attempting to understand the 'how' question, a lot of research has been focused on describing *what* people are doing. This research has led to identifying many behavioral patterns in people's probabilistic causal judgements, which are often described as systematic deviations of mean judgments from the CBN predictions (Rehder, 2014; Rottman & Hastie, 2014, 2016). Subsequently many explanations have been put forth to account for these behavioral patterns (Mistry et al., 2018; Rehder, 2014; Rottman & Hastie, 2016; Trueblood et al., 2017). These models have mostly been descriptive, and since they all target the same phenomena they can be hard to distinguish empirically. This difficulty is exacerbated by the fact that by and large the field has focused on the central tendency of responses, i.e. the mean. While focusing on the central tendency of responses is a principled approach and can be very effective, it leaves a lot of information contained in participant responses unused. With a target explanandum as rich as human causal cognition this has led to considerable difficulty in assessing the relative success of different candidate theories (Rehder, 2014, 2018; Rottman & Hastie, 2014, 2016).

One way out of this rut is to focus our efforts on analyzing distributions of causal judgments and not just their mean (Kolvoort, Temme, et al., 2023; O'Neill et al., 2022). While this comes with its own set of challenges, other fields have made large steps forward by doing just so. For instance, in the field of judgment and decision making the now widely-used evidence accumulation models (e.g. Ratcliff, 1978; Ratcliff et al., 2016), which account for the joint distribution of responses and response times, have allowed for important theoretical developments, such as an explanation of the speed-accuracy trade-off in decision-making (e.g. Bogacz, Wagenmakers, et al., 2010; Katsimpokis et al., 2020; Van Maanen et al., 2011), or an understanding of specific individual differences in decision-making behavior (Ratcliff et al., 2006; van Ravenzwaaij et al., 2011). In the field of judgment and decision-making, as well as others, it has been shown that the variability in behavior can be informative of the underlying cognitive processes and therefore its study can help constrain theoretical development. It has been known

[18] Also known as causal graphical models

for a while that there is a substantial amount of variability in causal probabilistic judgement with multiple authors commenting on this (Davis & Rehder, 2020; Kolvoort et al., 2021; Rehder, 2014; Rottman & Hastie, 2016). Hence we believe the time has come for the field of causal reasoning to exploit the variability in causal judgments and engage in the modelling of full response distributions (Kolvoort, Temme, et al., 2023).

Our main aim with the current study is to assess competing explanations for the variability in probabilistic causal judgements. To formalize this process and constrain theory development we will implement cognitive models of competing theories and subsequently perform model comparison. We will quantitatively assess the goodness-of-fit of the various proposed models to the data and perform model simulations to assess each model's ability to predict qualitative patterns (i.e. behavioral effects) of interest (Palminteri et al., 2017). As this approach requires a good estimate of the variability in judgments on the individual level, we re-analyze a dataset that contains both within- and between-participant variability (see Chapter 3; Kolvoort et al., 2021). In the next section we briefly discuss the dataset we will reanalyze in addition to providing an overview of important behavioral patterns in existing studies.

5.2 EXPERIMENTAL DATA

While modeling response distributions and targeting variability holds promise, it comes with its own challenges. One of the main difficulties lies with designing an experiment in such a way that it allows for the elicitation of repeated independent measurements. Repeated measurements are necessary to assess within-participant variability. Other fields that often elicit repeated measurements typically use stimuli material for which repeated presentation does not invoke practice effects (e.g. random-dot motion arrays). In the field of causal reasoning this is harder to do. Firstly, causal reasoning tends to require deliberative and conscious reasoning, making it more likely for participants to notice repetitions. Secondly, most causal judgment studies have stimuli consisting of states of causal variables, which are discrete symbols that could be memorized.

In recent experimental work we, for the first time, elicited repeated-measures of probabilistic causal judgments with the goal of assessing variability (Kolvoort et al., 2021).This dataset is suitable to compare the ability of different theoretical models to account for variability of judgments as it contains both within- and between-participant variability. We now provide a short description of the experiment and its findings (Kolvoort et al., 2021). For more details we refer the reader to the original paper.

5.2.1 Materials

Probabilistic causal judgments were tested in five domains: biology, sociology, astronomy, meteorology, and, sociology. Participants were first told that the causal network in the domain they were about to study had three binary variables. Next, they were presented with a verbal description of two causal relationships that formed a common cause network (Figure 5.1) where two variables were effects (henceforth X_1 and X_2) and one was the cause (Y). For each causal relationship a description was provided that included a discussion of the generative mechanism responsible for that relationship. All the causal variables and relationships were counterbalanced over participants. The domains and descriptions were all based on standard materials that have been used and validated by multiple other studies in the field (Rehder, 2014, 2018; Rehder & Waldmann, 2017).

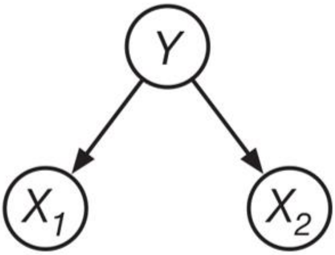
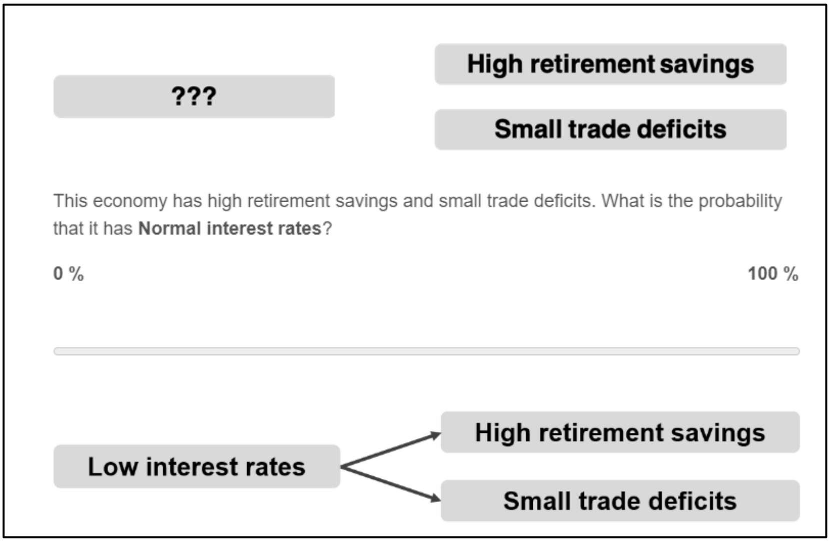


Figure 5.1 *Three-variable common cause network. Arrows denote causal relationships, circles denote causal variables.*

5.2.2 Procedure

After studying several screens with information about the overall task, for each domain participants were first presented with a cover story and a description of the domain's variables and causal relationships. The causal networks were described verbally and presented as a diagram. Participants were told that each variable's overall presence (i.e. the marginal probability) was 50%, and that each cause produced its effect "75% of the time". While a lot of information differed (e.g. the descriptions of the generative mechanisms), the underlying causal structure was the same for each domain. After learning this information about the domain participants were asked multiple comprehension check questions and could only continue after answering these correctly. Next was the inference test, in which each trial presented the values of one or two variables in the causal structure and asked participants to predict the state of another variable (Figure 5.2). Each domain consisted of 24 trials, and the order of trials and domains was randomized across participants. Participants responded by placing a tick on a rating scale ranging from 0% to 100%.



This economy has high retirement savings and small trade deficits. What is the probability that it has **Normal interest rates**?

Figure 5.2 *Screenshot of a trial in the experiment by (Kolvoort et al., 2021). This screenshot is of the* $P(Y = 1|X_i = 1, X_j = 1)$ *inference. Participants respond by clicking on the horizontal scale ranging from 0% to 100%. The bottom of the screen displayed the causal network on which participants were instructed to reduce memory load (see Rehder, 2018).*

5.2.3 Design and participants

Six different inference types were tested that varied on two factors: Information and Direction (Table 5.1). Direction referred to the direction of reasoning required, from cause to effect (Predictive) or from effect to cause (Diagnostic). Information refers variable values that are provided to the participants on each trial, which could either be Consistent (two variables with the same value), Inconsistent (two variables with differing values), and Incomplete (one variable). Within each domain participants responded to four different versions of each of the 6 inference types making 24 trials per domain. Different versions of each inference type were created by making use of the symmetric joint distribution that participants learned. For example, we could vary whether we asked about X_1 or X_2, and whether we referred to the presence or absence of a variable. As the joint distribution was symmetric in these terms we collapsed over these items to obtain a total of (4 versions x 5 domains =) 20 repeated measurements for each inference type.

		Reasoning Direction	
		Predictive	Diagnostic
Information	Consistent	$P(X_i = 1 \mid Y = 1, X_j = 1)$ = 80%	$P(Y = 1 \mid X_i = 1, X_j = 1)$ = 94%
	Incomplete	$P(X_i = 1 \mid Y = 1)$ = 80%	$P(Y = 1 \mid X_i = 1)$ = 80%
	Inconsistent	$P(X_i = 1 \mid Y = 1, X_j = 0)$ = 80%	$P(Y = 1 \mid X_i = 1, X_j = 0)$ = 50%

***Table 5.1** Inference types and normative answers for the experiment in Chapter 3 (Kolvoort et al., 2021). Inference types varied with two factors, Direction (predictive or diagnostic) and Information (consistent, incomplete, or inconsistent) resulting in 6 inference types. Xs and Ys refer to variables, where the Xs are effects, and the Y is the cause in a three-variable common cause network (see **Figure 5.1**). The 1s and 0s refer to the presence or absence of an effect or cause.*

5.2.4 Findings from model-free analyses

In the original analysis of the data we found that within-participant variability was lower for inferences with incomplete information, and that within-participant variability was higher for diagnostic inferences than for predictive ones (Kolvoort et al., 2021). Such systematic differences in variability support the premise that variability in causal judgments is informative of the underlying cognitive processes. In addition, we found some distinctive qualitative patterns in the response distributions of participants, that could be used to distinguish between theoretical proposals. Firstly, we found that a lot of response distributions were multimodal, often with one mode at 50%. Similar observations had been made in multiple probabilistic causal judgment studies (Kolvoort, Fisher, et al., 2023; Rehder, 2018; Rottman & Hastie, 2016). What was new was that the size of this mode at 50% seemed to vary as well, with relatively more judgments at (or close to) 50% when the information provided was less consistent (i.e. the mode was largest for Inconsistent information trials and smallest for Consistent information trials). In addition, the tendency to respond at 50% seemed larger for diagnostic inferences than for predictive ones.

The data from this study also displayed two hallmark features of human causal reasoning data. The first are Markov violations, which have been found in almost every experiment on causal reasoning (e.g. Ali et al., 2011; Davis & Rehder, 2020; Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Mayrhofer & Waldmann, 2015; Park & Sloman, 2013, 2014; Rehder, 2014, 2018; Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016; Sloman & Lagnado, 2015; Waldmann et al., 2008). Markov violations refer to the non-adherence to the Markov property of CBNs, which stipulates (conditional) independence between certain variables (see Rehder, 2018). In the repeated-measures experiment Markov independence relates to the predictive inferences (Table 5.1), where normatively the state of X_i should be independent of the state of X_j once the state of Y is known. We can state this formally as:

$$P(X_i = 1 \mid Y = 1, X_j = 1) = P(X_i = 1 \mid Y = 1) = P(X_i = 1 \mid Y = 1, X_j = 0)$$

However, people tend to judge that X_i is more likely to be present when X_j is also present, even when the state of Y is known. Hence people tend to judge that :

$$P(X_i = 1 \big| Y = 1, X_j = 1\big) > P(X_i = 1|Y = 1) > P\big(X_i = 1\big|Y = 1, X_j = 0\big)$$

A second hallmark feature of causal probabilistic reasoning is conservatism (e.g. Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Phillips & Edwards, 1966; Rottman & Hastie, 2014, 2016), which is also present in the dataset we analyze. This refers to the tendency of people to avoid the extreme ends of a response scale, and instead to respond more towards the middle of the scale (in this experiment this was at 50%). Moreover, it has been established that it is not just that the responses are conservative on average, but the actual bulk of people's individual responses are conservative, meaning they fall between 50% and the normative response (Kolvoort, Fisher, et al., 2023; Kolvoort, Temme, et al., 2023; Rottman & Hastie, 2016).

Table 5.2 provides a list of the qualitative patterns in the data from previous (probabilistic) causal judgement experiments with references to studies where they were observed. A complete mechanistic theory of causal reasoning should be able to explain these patterns.

We already made a first step towards using modeling to explain the variability of causal judgements and certain qualitative patterns in a previous study (Kolvoort, Temme, et al., 2023). It was the first study in this field that showcased the possibility to fit models which predict variability to raw response data and to target distributional phenomena in those responses. However, that study had two main limitations. The first limitation was that the data on which the models were compared did not include repeated measurements (see Kolvoort, Fisher, et al., 2023), limiting the conclusions we could draw. Second, the study only tested the Mutation Sampler (MS; Davis & Rehder, 2020) and a generalization we developed called the Bayesian Mutation Sampler (BMS; Kolvoort, Temme, et al., 2023). While the BMS and MS are promising theoretical proposals for the cognitive mechanisms underlying causal judgments, other plausible theories and sources of variability exist that could potentially explain distributions of causal judgments as well. In the current study we will include all these other candidate theories and we will now discuss them in more detail.

Qualitative pattern	Explanation	Studies
1a. Mean conservatism	Mean response tend to be between normative probability and 50%	(Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Phillips & Edwards, 1966; Rottman & Hastie, 2014, 2016)
1b. 'moderate' conservatism	Bulk of responses lie between normative probability and 50%	(Kolvoort et al., 2021; Kolvoort, Temme, et al., 2023; Rottman & Hastie, 2016)
1c. Extreme responses are rare	Participants tend to avoid the extremes of the response scale, in probabilistic causal judgement tasks this is near 0% and 100%	(Davis & Rehder, 2020; Kolvoort, Temme, et al., 2023; Rottman & Hastie, 2016)
2. Markov violations	Non-adherence to Markov property, which refers to the (conditional) independence of causal variables. In the case of a common cause network, this is the independence of X_i and X_j once the state of Y is known. This phenomenon is also referred to as 'failures to screen off'	(Ali et al., 2011; Davis & Rehder, 2020; Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Mayrhofer & Waldmann, 2015; Park & Sloman, 2013, 2014; Rehder, 2014, 2018; Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016; Sloman & Lagnado, 2015; Waldmann et al., 2008)
3a. Within-participant variability is lower for incomplete information	Responses to queries with incomplete information are less variable	(Kolvoort et al., 2021)
3b. Within-participant variability is higher for diagnostic inferences	Judgements are more variable when participants are asked to reason from effect to cause (Diagnostic) as compared to when they reason from cause to effect (Predictive)	(Kolvoort et al., 2021)
4. Multi-modal response distributions	Response distributions often have more than one mode, even on the participant level	(Davis & Rehder, 2020; Kolvoort et al., 2021; Kolvoort, Temme, et al., 2023; Rottman & Hastie, 2016)
5a. Spikes at 50%	Response distributions often have a mode or 'spike' of responses at 50%	(Davis & Rehder, 2020; Kolvoort et al., 2021; Kolvoort, Fisher, et al., 2023; Rehder, 2018; Rottman & Hastie, 2016)
5b. Spikes at 50% increase with inconsistency of information provided	Participants tend to respond at 50% less when consistent information is provided, and more when inconsistent information is provided, compared to when the information is incomplete	(Kolvoort et al., 2021; Rottman & Hastie, 2016)
5c. Spikes at 50% are larger for diagnostic inferences	Participants tend to respond at 50% more so when they are asked to reason from effect to cause (Diagnostic) as compared to when they reason from cause to effect (Predictive)	(Kolvoort et al., 2021)

Table 5.2 *Overview of qualitative patterns in response distributions of causal probabilistic judgments.*

5.3 CANDIDATE MODELS

We identified two models in the causal reasoning literature that predict within-participant variability in causal judgments. These are the Bayesian Mutation Sampler (Kolvoort, Temme, et al., 2023) and the Beta Inference Model (Rottman & Hastie, 2016). In addition to these models from the literature, we developed four simple models based on general psychological mechanisms that could possibly introduce the type of variability in responses that has been observed. These we have termed the Motor Variability Model, the Stimulus Encoding Error model, the Parameter Uncertainty Model, and the Guessing model. We will now give an overview of all these models and discuss their psychological justification as well as the mathematical implementation that we will use for comparing these models.

5.3.1 Bayesian Mutation Sampler

The Bayesian Mutation Sampler (BMS; Kolvoort, Temme, et al., 2023) is a generalization of the Mutation Sampler (MS; Davis & Rehder, 2017, 2020; Rehder & Davis, 2021), and developed to account for distributions of causal judgments. The MS is a model of causal reasoning that assumes that individuals draw resource-constrained inferences based on a sampling process of CBNs. The model proposes that people think of concrete cases when asked to reason about a causal system. These concrete cases are causal systems where each variable is instantiated with a value (e.g. [X_1 = 1, Y = 1, X_2 = 0]) and they are retrieved from memory or generated using an internal generative model. The MS assumes people sample these cases using a Metropolis-Hastings algorithm, which is a Markov chain Monte Carlo sampling method for approximating probability distributions (Hastings, 1970; van Ravenzwaaij et al., 2018). This method converges to the true distribution when the number of samples grows large. The MS, however, assumes that people are restricted in the number of samples they can take as they are restricted in cognitive resources. This limited sampling in combination with two other assumptions is what makes in the MS accurate in predicting mean responses on causal judgement tasks (Davis & Rehder, 2017, 2020; Rehder & Davis, 2021). The first assumption is that the proposal distribution for each step of sampling consists only of those states that differ from the current state by only one variable (i.e. the current state is 'mutated'). This assumption implements the idea that when people think of a next case, that case is likely similar to the one they are thinking of currently. The second additional assumption is that the sampling process starts out at a prototypical state, which is a state in which all causal variables are either present or absent (in our case either [X_1 = 0, Y = 0, X_2 = 0] or [X_1 = 1, Y = 1, X_2 = 1]), as these states readily come to mind. Because the MS posits that people only take a limited number of samples, the proposal distribution and starting point bias the sampling process such that approximated distribution assigns more (less) probability weight to states with consistent (inconsistent) variable values than the normative distribution (see for more details Davis & Rehder, 2020). The size of this bias depends on the number of samples, i.e. the chain length, which is a free parameter of the model.

After generating a chain of samples, the relative frequencies of the obtained samples in the chain are used to estimate the probability query. For example, if we obtained the two identical samples [A = 1, B=1] and [A=1, B=1], we would estimate $P(A = 1 \,|B = 1)$ as 100%, since in our set of samples in all cases where B = 1 we have that A = 1. However, it can be the case that our samples do not contain the right states to compute the right relative frequency. For example, if we would want to estimate $P(A = 1 \,|B = 1)$, but all the states in our chain of samples have that B = 0. In such a case the MS defaults to responding with 50%.

While the MS has been shown to be able to predict mean responses (Davis & Rehder, 2020; Rehder & Davis, 2021), it failed to predict to observed patterns of response distributions (Kolvoort, Temme, et al., 2023). The BMS, instead of computing judgments directly from frequencies in the obtained samples, combines the information gained from sampling with generic prior information to generate a judgment. The integration of prior information yields a better explanation of response distributions (Kolvoort, Temme, et al., 2023). The BMS has two free parameters, the chain length and the β prior parameter that determines the shape of the symmetric Beta distribution used as a prior. We implemented only the BMS as it generalizes the MS, i.e. when the β prior parameter of the BMS is 0 the model is equivalent to the MS.

5.3.2 Beta Inference model

Rottman and Hastie (2016) proposed a model of causal inference called the Beta Inference Model (BIM) to explain Markov violations and variability in judgments. The motivation for the BIM was that when in an experiment participants are asked to learn about a causal system by experience (i.e. by viewing samples of data from the causal system, not reading descriptions of the causal system as in the repeated-measures experiment), then it is possible to compute judgments directly from the samples of data provided for learning. The BIM considers an inference such as $P(X_i = 1|Y = 1)$, to be a problem of computing the proportion of times that $X_j = 1$ ("win") versus $X_j = 0$ ("failure") within the set of cases where Y = 1. The posterior distribution of that proportion is then given by a Beta distribution which describes participant judgments (Rottman & Hastie, 2016). In this sense the model proposes that people infer a posterior distribution directly from the data they viewed to learn about the causal system (by regarding the "wins" and "failures" in the learning data), and then sample from this distribution when making the inference (Rottman & Hastie, 2016). As the learning data provided to participants represents underlying joint distributions truthfully, the modes of the predicted distributions coincide with the CBN point predictions. However, the skewness and concentration of the predicted Beta distribution changes depending on the amount of learning data which is directly related to the conditional statement, in our example Y = 1. Differences in skewness (due to differences in the conditional statement) can for example explain Markov violations as the mean of responses can shift away from the mode (which is fixed at the location of the normative probability).

As discussed, the BIM as originally proposed assumes that individuals learn about a causal structure by viewing data. However, the behavioral features it accounts for, such as Markov violations, have been found in experiments that do not use such a learning-by-experience procedure (Rehder & Waldmann, 2017). The BIM can be implemented without learning data by assuming that samples come from an internal generative model (cf. Rottman & Hastie, 2016), similar to the BMS. Therefore, in the absence of a learning-by-experience procedure in the repeated-measures data we model, we treat the number of samples as a free parameter in the model, instead of inferring the number of samples from the experimental design. However, the ratio between the number of samples for each inference can be inferred from the experimental design, as this is determined by the joint distribution of the causal network that participants learn.

For instance, the probability that the conditional statement in $P(X_i = 1|Y = 1)$ is satisfied, is higher than the probability that the conditional statement in $P\left(X_i = 1\middle|Y = 1, X_j = 1\right)$ is satisfied, as $P(Y = 1) > P\left(Y = 1, X_j = 1\right)$. Thus, there will be more samples for the former inference than for the latter, determined by the ratio of $P(Y = 1)$ to $P(Y = 1, X_j = 1)$ (Table 5.3). That there are more samples available for the former leads the BIM to predict a more concentrated distribution of responses for that inference. This generalized version of the BIM assumes that people generate a varying number of samples for different inference types in particular ratios via their generative model[19]. We could theorize that this is due to the states with higher/lower probability being harder/easier to generate, which is analogous to what is proposed by the (B)MS to govern sample generation (cf. Davis & Rehder, 2020; Kolvoort, Temme, et al., 2023).

This generalized BIM predicts responses to be drawn from Beta distributions, where the modes are centered on the CBN point prediction, and where the concentration of the Beta distribution is set by the amount of learning samples. That is, if there are more samples, available responses fall nearer to the CBN point prediction. Hence, for different inference types we can scale the concentration as determined by the conditional statement. The concentration scaling factors for the current repeated measures experiment are presented in Table 5.3. From this table we can see that the BIM predicts response to the $P\left(X_i = 1\middle|Y = 1, X_j = 0\right)$ to be most variable, while responses to the $P\left(Y = 1\middle|X_j = 1\right)$ and $P(X_i = 1|Y = 1)$ inferences are predicted to be least variable. The BIM has only one free parameter, the concentration of the predicted Beta distribution for the $P\left(X_i = 1\middle|Y = 1, X_j = 0\right)$ inference. This concentration parameter is theoretically equivalent to the amount of generated samples for that inference. The concentration of the Beta distributions for the other inference types is derived by multiplying this base concentration by the concentration scaling factors in Table 5.3.

Concentration scaling for Beta Inference model

Inference	Total samples in possible learning data	Probability of sampling from generative model	Concentration scaling factor
$P(X_i = 1\|Y = 1)$	$N(Y = 1)$	$P(Y = 1) = .5$	5
$P\left(X_i = 1\middle\|Y = 1, X_j = 0\right)$	$N(Y = 1, X_j = 0)$	$P\left(Y = 1, X_j = 0\right) = .1$	1
$P\left(X_i = 1\middle\|Y = 1, X_j = 1\right)$	$N(Y = 1, X_j = 1)$	$P\left(Y = 1, X_j = 1\right) = .4$	4
$P\left(Y = 1\middle\|X_j = 1\right)$	$N(X_j = 1)$	$P\left(X_j = 1\right) = .5$	5
$P\left(Y = 1\middle\|X_i = 1, X_j = 0\right)$	$N(X_i = 1, X_j = 0)$	$P(X_i = 1, X_j = 0) = .16$	1.6
$P\left(Y = 1\middle\|X_i = 1, X_j = 1\right)$	$N(X_i = 1, X_j = 1)$	$P\left(X_i = 1, X_j = 1\right) = .34$	3.4

Table 5.3 *Concentration scaling for Beta Inference Model (BIM). These scaling factors determine the relative concentrations of the predicted response distributions. The probabilities and scaling factors are derived from the parametrization of the causal network taught to participants in the repeated-measures experiment (Kolvoort et al., 2021).*

[19] If, instead, we would assume that people generate an equal number of samples for each inference then the model would be equivalent to the Motor variability model discussed below.

5.3.3 Motor Variability model

In addition to the BMS and BIM that have been previously proposed, we implemented four new models to account for variability in causal judgments. The first of these models posits that people reason normatively, but their responses vary from trial to trial due to motor noise (or general task noise). In other domains, there is ample evidence that motor noise introduces variability in judgments (e.g., Maaß et al., 2021; Müller & Sternad, 2004; Verdonck & Tuerlinckx, 2013). The idea that people reason normatively to some extent but that such reasoning interacts with 'non-normative' processes to result in judgments has been proposed before (see Rehder & Waldmann, 2017; Rottman & Hastie, 2016).

According to this motor noise explanation we should expect a distribution of responses where the mode is centered at the normative probability. As responses are restricted between probabilities of 0 and 1, a natural way to model this is using the Beta distribution. Note that this model is closely related to the BIM, where the distributions are also centered on the normative response. The main difference with the Beta Inference model is that for the Motor Variability Model (MVM) the variability is the same for each inference type, as the amount of motor noise should not be affected by the content of a stimulus. The MVM has only one free parameter, the concentration parameter of the noise distribution.

5.3.4 Stimulus encoding error model

The second model we developed here is based on the idea that participants can misread part of the stimulus, and sometimes making such an error results in reading off a different variable value than presented. For instance, instead of correctly encoding the stimulus equivalent of $P\left(X_i = 1\middle|Y = 0, X_j = 1\right)$, a participant could erroneously encode the stimulus as $P\left(X_i = 1\middle|Y = 1, X_j = 0\right)$. We refer to this as the Stimulus Encoding Error (SEE) model. One reason to include such a model in our analysis is that it has been found previously that individual-level response distributions are multimodal (Kolvoort et al., 2021). Misreading variable values on some trials, but not others, is a simple mechanism which would lead to multimodal distributions.

The SEE model assumes that a participant misreads the state of a conditioning variable with probability m. For inferences with one conditioning variable (e.g. $P(X_i = 1|Y = 1)$) this means that the model predicts a probability of 1-m correct responses, and a probability of m incorrect responses (i.e. responses to $P(X_i = 1|Y = 0)$). For inferences with two conditioning variables (e.g. $P\left(X_i = 1\middle|Y = 1, X_j = 1\right)$) we still have that m is the probability of independently misreading the value of each conditioning variable. This leads to the probability of misreading the values of both variables being m^2. The probability of only the first variable being misread is the same as the probability of only the second being misread, namely $m - m^2$. The probability of neither being misread is therefore $1 - 2m + m^2$. As misreading a variable value can lead to responses very different from the normative response this mechanism can predict multiple modes (as can been seen from the number of different possible responses per row in Table 5.4). Using this implementation, the SEE model has only one free parameter, the probability m of misreading the state of a conditioning variable.

Predictions of SEE model

Inference	Normative prob. (misreading neither)	Possible 'erroneous' responses			Possible # of modes when m>0
		misread first	misread second	misread both	
$P(X_i = 1 \mid Y = 1)$	.8	.2	-	-	2
$P(X_i = 1 \mid Y = 1, X_j = 0)$	.8	.2	.8	.2	2
$P(X_i = 1 \mid Y = 1, X_j = 1)$	.8	.2	.8	.2	2
$P(Y = 1 \mid X_i = 1)$	.8	.2	-	-	2
$P(Y = 1 \mid X_i = 1, X_j = 0)$	.5	.06	.94	.5	3
$P(Y = 1 \mid X_i = 1, X_j = 1)$	.94	.5	.5	.06	3

Table 5.4 *Predictions of Stimulus Encoding Error (SEE) model. The free parameter m is the probability of misreading the state of a conditional variable. Probability of only misreading the first conditional variable state is m-m^2, misreading second is also m-m^2, and misreading both is m^2. The probability of misreading neither is* $1 - 2m + m^2$.

5.3.5 Parameter Uncertainty model

Another possible source of (within-participant) variability is uncertainty regarding the causal parameters, i.e. the base rates[20] and causal strength parameters. We refer to this account as the Parameter Uncertainty Model (PUM). Such uncertainty could lead a reasoner to use slightly different values for these parameters every time they are used to compute a response. Because – according to CBN – not all causal parameters are necessary to compute every inference, the amount of variability the PUM predicts varies per inference type. For instance, diagnostic inferences require information regarding the strength of background causes for both the cause and effect, while predictive inferences only require information regarding the background causes of the effect.

We model parameter uncertainty by first drawing the causal parameters (base rates, causal strengths) from a Beta distribution centered on the normative parameter value. Next, these noisy causal parameters are used to compute a judgement according to the normative CBN framework. While the Beta distribution has its mode at the normative probability of the parameter, the concentration of the distribution can vary as a free parameter to model different levels of uncertainty. Moreover, as the causal parameters are of two types, base rates and causal strengths, each type has its own concentration parameter to model possible reasoners that are more uncertain about one type than the other. Consequently, the PUM has two free parameters, namely the concentration parameters for the Beta distributions from which the base rates and causal strengths will be sampled.

[20] We use the term 'base rate' to refer to the probability of a causal variable being present without its causes in the modeled causal structure being present. This is sometimes referred to as the "strength of alternative causes".

5.3.6 Mixture modeling using a guess component

One thing to note from the above discussion of models is that the SEE model, the MVM, and the PUM cannot produce Markov violations on their own. These models predict the same distribution of responses for each of the predictive inference types known to exhibit Markov violations. This fact makes these models implausible on their own, as Markov violations are often claimed to be a hallmark of human causal reasoning and so we should expect distributions of these judgments to differ at least in their mean. However, it is plausible that on some trials participants respond using a different generative mechanism than specified by the above models. This additional process could then produce the Markov violations that have been observed.

This reasoning is part of our motivation for the last source of variability we model here with the Guess Model. In addition, this Guess Model is inspired by the observation that response distributions tend to feature a large spike of responses at 50% (Kolvoort et al., 2021; Kolvoort, Temme, et al., 2023; Rottman & Hastie, 2016). One possible explanation for this observation is that participants, instead of generating a response according to any of the theories explained so far, simply respond with 50% as a type of default response or guess (Rottman & Hastie, 2016). There is some evidence for this conjecture as it has been found that participants have less confidence in responses near 50% (Kolvoort, Fisher, et al., 2023), suggesting that they did not know the answer. Such guessing or default responding could explain average Markov violations if the amount of guessing is dependent on the inference type.

There are many possible reasons for why a participant would guess, such as a lack of task compliance, a lapse in concentration, being distracted, or not understanding the stimulus (e.g. in case of inconsistent trials). Whatever the underlying reason may be, it seems reasonable that on some trials participants just respond with the middle of the scale as a default response option, regardless of what the judgment generating process is on the other trials. The Guess Model thus adds a guess component to all the models discussed above. This makes the resultant predictions mixtures of a guessing and reasoning component, with the latter provided by the five base models (BMS, BIM, MVM, SEE, and PUM).

The Guess Model stipulates a probability of a participant to respond at 50% by default rather than computing an answer. As previously summarized, spikes at 50% have been found to be largest for inconsistent and for diagnostic inferences (Kolvoort et al., 2021). It has been noted before that people tend to find diagnostic reasoning more difficult than predictive reasoning (Fernbach et al., 2011; Fernbach & Darlow, 2010). Similarly, it seems inconsistent inferences can be more difficult as the stimulus provides conflicting cues. Based on this we hypothesize that guessing might be more likely for problems that are perceived to be harder or more complex by participants. That is, the experimental factors Information and Direction might affect how often participants guess, and so we implement five different versions of this model with different constraints. In the simplest version (0 parameters), there is no guessing. Additionally, we implemented models in which participants guess with a fixed proportion over all inferences (1 free parameter), in which they change in their guessing based on reasoning Direction (2 parameters), in which they change their guessing based on the Information provided (3 parameters), and in which guessing changes based on both Direction and Information, leading to different guess proportions for each inference type (6 parameters).

5.4 Fitting and Analysis Methodology

5.4.1 Fitting Procedure

To fit the theoretical models described above to the data we use a simulation-based approach combined with an exhaustive grid search. This is a form of 'pre-paid' estimation (Mestdagh et al., 2019), which falls under the umbrella of amortized inference methods (Radev et al., 2020). We use such a grid search as it removes possible bias in the optimization procedure. Such a bias has been found to be severe for fitting the Mutation Sampler and Bayesian Mutation Sampler models with a step-wise optimization procedure (Kolvoort, Temme, et al., 2023). We will fit the models to each participant separately. As we keep parameters fixed for the base models over the different inference types in the experiment, this means that we will fit each set of parameters to (6 inference types x 20 repeated measurements =) 120 responses.

In the first step of the fitting procedure, we simulate responses using the models and save these simulated responses in a grid. We choose a range of realistic parameters (see below) for these simulations so that the grid covers plausible response distributions under the different models. We will simulate 100,000 responses for every combination of parameters for each model. Next, we use Probability Density Approximation (PDA; Holmes, 2015; Turner & Sederberg, 2014) to construct a 'synthetic' likelihood for each cell in the grid by way of kernel density estimation. This provides a likelihood of observing the data under each model and set of parameters, for each cell of the grid. The best fitting model and parameters for each participant are then given by the cell with the highest likelihood. One important parameter setting for the PDA method is the kernel bandwidth (see Lin et al., 2019). We picked our bandwidth on the basis of the dataset, so that the predicted distributions will match the granularity of responses. To do this, we applied the Sheather and Jones method for non-parametric automatic bandwidth selection (Jones et al., 1996; Sheather & Jones, 1991) to each participant's data and averaged those. This average was 2.13 (on percentage point scale), which we used as the standard deviation of a gaussian kernel for the kernel density estimation (cf. Lin et al., 2019). In the next step, we compute BIC values and weights (Schwarz, 1978; Wagenmakers & Farrell, 2004) to compare the fit of the models for each participant.

5.4.2 Considerations for determining the grid

We aimed to have 20 values[21] for each free parameter of the main models and 7 values for the Guessing model component (see below), to cover the range of plausible values. This number of parameter values is restricted due to computational resources related to the combination of the main models with the guess proportions. For the BMS, previous work has indicated that it is more important to have a non-biased optimization procedure than to have a very precise parameter estimate as small differences in parameter estimates only lead to small differences in the predictions (see Appendix A in Kolvoort, Temme, et al., 2023).

[21] For the β prior parameter of the BMS and for the probability of misreading a stimulus value *m* of the SEE model we use 21 values as they naturally lend themselves to picking an odd number of values, see below.

The BIM, MVM, and PUM make use of Beta distributions in their implementation. For these Beta distributions we do not use the standard parametrization in terms of shape parameters, but rather use an alternative parametrization in terms of the concentration and mode to define the distributions (N. L. Johnson et al., 1995). For these models the concentration is a free parameter for which we will pick grid values. The other parameter, the location of the mode, is fixed by the model specification (at the normative response for the BIM and MVM, and at the normative causal strength and base rate values for the PUM).

The BMS and PUM have two free parameters and the BIM, MVM and SEE have one free parameter. Additionally, the guess model has a maximum of 6 parameters (one guess proportion for each of 6 inference types when it is allowed to vary for both Direction and Information factors). Following this, we have for the BMS and Parameter Variability models a grid with at least (20 x 20 x 7 x 7 x 7 x 7 x 7 x 7 =) 47,059,600 unique parameter combinations, and for the other models we have grids with (20 x 7 x 7 x 7 x 7 x 7 x 7 =) 2,352,980 at least parameter combinations.

5.4.2.1 Bayesian Mutation Sampler

The chain length of the BMS is estimated between 6 and 200. This includes the range of chain lengths found for causal reasoning tasks before (Davis & Rehder, 2020). When chain lengths become large the differences in predictions become negligible (Davis & Rehder, 2020). Therefore, we used equally spaced points between 6 and 200 in terms of their inverse cube root. That is, we computed the inverse cube root ($x^{-\frac{1}{3}}$) of 6 and 200, then picked 18 equally spaced points between them (so in total we have 20 values including $6^{-\frac{1}{3}}$ and $200^{-\frac{1}{3}}$) and then reverted all these points back to the original space and rounding them, resulting in the following sequence of chain lengths: [6, 7, 8, 9, 11, 12, 14, 16, 18, 21, 25, 30, 36, 43, 53, 66, 84, 110, 146, 200].

For the β parameter of the Beta(β, β) prior we wanted the grid values to be picked such that the prior distributions were symmetric around the uniform distribution, which occurs when β = 1. To do this we first picked values for β < 1 from 0 to 1 with a step size of 0.1. Next, to pick values for β > 1, we computed the total variation distance (Levin & Peres, 2017) between the uniform distribution and each Beta(β, β) distribution with β < 1, and then identified a set of β > 1 that have the same total variation distances from the uniform distribution (Kolvoort, Temme, et al., 2023). This symmetry about the uniform prior was not used for picking a value symmetric to β = 0, as this would lead to an infinite value, instead we use a value of 100. This procedure gives the following 21 values for β: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.11, 1.26, 1.45, 1.73, 2.14, 2.83, 4.14, 7.35, 21.54, and 100].

5.4.2.2 Beta Inference Model

The concentration of the Beta distributions predicted as response distributions is the only free parameter for the BIM. The grid values we pick will represent the concentration of the predicted response distribution for the $P(X_i = 1 | Y = 1, X_j = 0)$ inference, which is the inference which has the lowest concentration (see Table 5.3). The concentrations of the distributions for the other inference types is determined by multiplying the concentration with the respective concentration scaling factor. At the lower limit of the set of concentrations we want the model to predict a uniform distribution, while at the upper limit we want the model to predict no variability, which is for responses to all be at a single percentage point. As the lower limit we use a concentration of 2, which results in a uniform distribution. Moreover, since for the Beta Inference model the concentration is theoretically equivalent to the number of learning experiences or internally generated samples, 2 is a practical lower limit as a value of 1 would result in only 0% or 100% responses. To determine the upper limit, we identified 2^13=8192 to be the lowest power of 2 for which the standard deviation of the Beta distribution would be within half a percentage point, meaning that the majority of probability weight would be assigned to a single percentage point. Consequently, the grid range for the concentration parameter is [2^1, 2^{13}]. As with the chain lengths for the BMS, the difference between the predictions becomes smaller for larger concentrations (Figure 5.3). Therefore, we applied the same inverse cube root procedure as for the BMS chain lengths, resulting in the following set of concentration values: [2, 3, 4, 5, 6, 7, 9, 11, 13, 17, 23, 30, 43, 62, 96, 158, 290, 620, 1722, 8192].

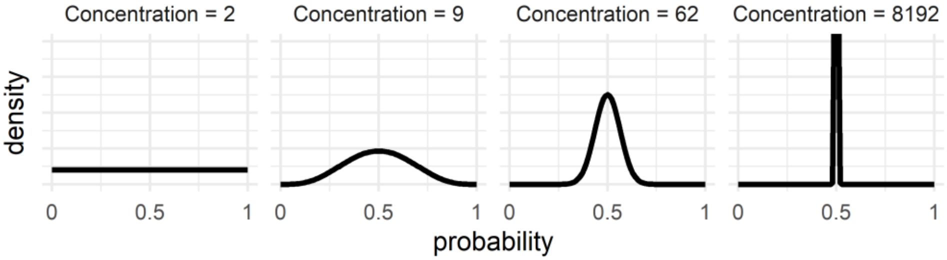


Figure 5.3 *Beta distributions, centered on .5, with a range of concentrations used in the grid search for the Beta Inference model, the Motor Variability model, and the Parameter Uncertainty model. The concentrations plotted here are equally spaced in the set of concentrations used in the grid.*

5.4.2.3 Motor Variability Model

The MVM has concentration as the only free parameter as well. The same reasoning as for the Beta Inference model applies here, so we use the same set of concentration values as for the Beta Inference model.

5.4.2.4 Stimulus Encoding Error model

For the SEE model the probability of misreading a stimulus value is the single free parameter. This ranges naturally from 0 to 1, and we picked 19 equally spaced values resulting in the following set of 21 values: [0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1].

5.4.2.5 Parameter Uncertainty Model

The PUM has two free parameters: the concentrations of the Beta distributions from which the causal strengths and base rates are sampled. The mode of these Beta distributions is fixed at the true value of the parameter (.5 for the causal strengths, .5 for the base rate of the cause (Y), and .2 for the base rates of the two effects (X_1 and X_2), see discussion of the data above). The estimation of the concentration of these distributions is the same for the BIM and MVM models.

5.4.2.6 Guess Model

For the Guessing model, we are restricted to a coarser grid due to computational resource restrictions which stem from the fact that this model is combined with all the aforementioned models. However, by looking at the individual response distributions for each inference type (Kolvoort et al., 2021, fig. 1) we can pick these values in a way that covers plausible values. The first thing to note is that some distributions have all responses at 50% and some have none, so we need to include 0 and 1 as guess proportions. Next, we can observe that in the case of bimodal response distributions, the mode at 50% is never larger than the other mode, meaning that at least half of responses were not at 50%. This leaves the range between 0 and .5, and we picked values with step size 0.1 between these limits, resulting in the following set of 7 guess proportions: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 1]. As participants provide 20 responses to each inference type, these parameter values correspond to 0, 2, 4, 8, 10, or 20 guesses respectively.

5.4.3 Analyses

For our analysis of model predictions and parameters we will make use of Bayesian Model-Averaging (BMA; Hinne et al., 2020; Hoeting et al., 1999). BMA allows for inference regarding model parameters and predictions while taking into account the uncertainty regarding the best model. It does this by assigning weights to each model based on the posterior model probabilities. These will be computed by comparing the relative Bayesian Information Criterion (BIC) scores (Schwarz, 1978; Wagenmakers & Farrell, 2004). We will use BMA within base models, i.e. we collapse over the varying constraints used for the Guess Model component, to analyze the base model parameters and predictions. Additionally, we will collapse over all models to obtain BMA-weighted estimates of the guess proportions. To compare parameter estimates between groups we will use the non-parametric Mann-Whitney-Wilcoxon test (Mann & Whitney, 1957; van Doorn et al., 2020) instead of the standard t-test, as parameter values are not normally distributed (see previous section).

We will analyze the predicted effects and parameters using repeated-measures ANOVAs, computing Bayes Factors using the BayesFactor package in R (Morey & Rouder, 2014). For any post-hoc comparisons we will use Tukey’s HSD to correct p-values for multiple comparisons. As an index of variability we will use Gini’s Mean Difference (GMD; David, 1968; Yitzhaki, 2003), defined as the average distance between any two observations. We use GMD instead of the parametric standard deviation as responses tend not to be normally distributed.

5.5 RESULTS

5.5.1 Overall model fit

Our main goal is to identify the most likely theoretical model considering the distributions of causal judgments, and therefore we begin by examining the quantitative fit of each model. To assess relative model fit we computed BIC weights of all models. We did this at the group and individual levels. We find that the BMS is the most likely model at the group level (Table 5.5).

Aggregating over guess components and computing group-level posterior model weights (Wagenmakers & Farrell, 2004), we find that the BMS is more than 1,000 times more likely to be the true data-generating process compared to any of the other models (evidence ratios BMS: 10^{293} versus BIM, 10^{120} versus MVM, 10^{57} versus PUM, and larger than 10^{300} versus SEE). This indicates that the stochastic sampling mechanism underlying the BMS is likely to be an important source of variability in causal judgments.

We find that the BIC score of each base model, except SEE, at the group level is best when a guess component with 6 parameters is added (group BIC columns in Table 5.5), reflecting that each model requires different guess proportions for each inference type to account for the whole dataset. However, this is not the case at the individual level. The BMS without guessing is the best model for most participants. And for each of the base models the added complexity of the guess parameters does not always improve the balance between fit and model complexity (individual BIC columns in Table 5.5), indicating that participants vary in their guess proportions. This suggests that there are relevant individual differences and raises the question of whether BMS is the most likely model for every participant. To answer this question we computed posterior model probabilities (based on BIC weights) for each model separately for each participant (Figure **5.4**).

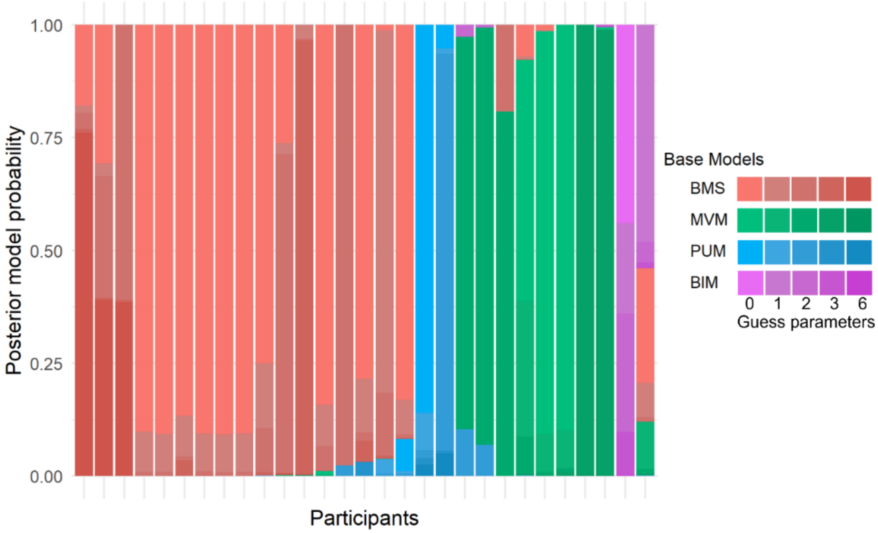


Figure 5.4 *Posterior model probabilities, based on BIC weights, for each participant. The darker the particular colors the more guess parameters are estimated for that model. 0 parameters refers to no guessing, 1 parameter to a fixed amount of guessing for all inference types, 2 parameters refers to different guess proportions per reasoning direction, 3 parameters refers to different guess proportions per type of information provided, and 6 parameters refers to a different guess proportion per inference type.*

		BMS			PUM			BIM			MVM			SEE		
		Group BIC	Ind. BIC	Win	Group BIC	Ind. BIC	Win	Group BIC	Ind. BIC	Win	Group BIC	Ind. BIC	Win	Group BIC	Ind. BIC	Win
No guessing		-3737	-120.1 (12.8)	10	-2504	-77.53 (12.3)	1	-1154	-35.40 (17.2)	1	-2403	-78.47 (11.5)	3	47157	1630 (170)	0
Guessing component	1 par	-4097	-128.1 (14.3)	1	-3660	-113.0 (14.6)	0	-2540	-78.78 (20.1)	1	-3427	-109.4 (15.0)	0	29751	1035 (145)	0
	2 par	-4432	-135.2 (16.9)	3	-4049	-122.0 (17.9)	1	-2972	-89.29 (23.3)	0	-3811	-118.2 (18.5)	3	29361	1026 (147)	0
	3 par	-4277	-125.5 (15.3)	1	-3969	-114.9 (16.0)	0	-2831	-80.02 (21.6)	0	-3676	-109.2 (16.6)	2	29471	1034 (145)	0
	6 par	-4679	-126.2 (17.8)	1	-4415	-117.0 (16.9)	0	-3327	-83.92 (24.4)	0	-4127	-111.5 (19.7)	0	29042	1032 (147)	0
Overall			-143.6 (17.4)	17		-131.2 (18.6)	2		-97.84 (24.0)	2		-127.4 (19.3)	8		1021 (147)	0

***Table 5.5** Wins, Group BICs, Mean individual BICs, and their standard errors per model. Group BIC refers to group-level BIC values computed using all data. Ind. BIC refers to individual-level BIC values averaged over participants to get the mean BIC values and the standard error (in brackets). For the 'Overall' row these are the average of the best fitting models for each participant. Lower BIC values imply a better fit. The win column refers to the number of participants for which the model is the winning model (out of 29 participants). The guess components are listed per number of parameters, where 1 par refers to a fixed amount of guessing for all inference types, 2 par refers to different guess proportions per reasoning direction, 3 par refers to different guess proportions per type of information provided, and 6 par refers to a different guess proportion per inference type.*

We find that the majority of participants (17 participants, 59%) are best explained by the BMS (Figure 5.4). This provides additional evidence suggesting that the sampling scheme of the BMS is an important source of variability, but not necessarily for every participant. This opens the door for the idea that the dominant sources of variability can vary per participant. This is evidenced by the finding that for each participant, except for one (the rightmost column in Figure 5.4), there is a base model that clearly explains their data best. For 2 (7%) participants the PUM model is preferred, for 8 (28%) participants the MVM is best, and for 1 (3%) participant we have that the BIM outperforms the other models. The SEE model, on the other hand, is outperformed by the other models for every participant, and so the misreading of stimulus values is an unlikely source of variability. All the other models are at least best for some participants, but sampling (BMS) and motor variability (MVM) seem the most probable sources of variability, as their respective models are the winning models for sizable groups of participants. This raises the question why these two groups of participants differ in their best fitting model. It could be that it is due to these groups having different response strategies, or possibly due to the statistical properties of the models. To investigate this further and to assess whether the models can capture the relevant behavioral effects we look at the model predictions and pay specific attention to the patterns in the response data that we identified in Section 5.2.

5.5.2 Predicted Means and Variability

To assess predictive model performance we computed model-averaged predictions for each of the base models. We then computed the mean and GMD predictions of each base model for each participant and inference type so that we can analyze the predicted patterns in central tendency and variability respectively. Overall, the differences between observed and predicted means and GMD are relatively small, with each model's predictions being on average around 5 to 6 percentage points off from the observed responses (Table 5.6). Using repeated-measures ANOVAs, we find that the models overall do not differ in terms of how well they predict the mean response ($F(4, 837) = 0.967$, $p = .425$, $\text{BF}_{10} = 0.0061$), but we do find evidence that the models differ in how well they predict GMD ($F(4, 837) = 4.89$, $p < .001$, $\text{BF}_{10} = 2.31$). Using post-hoc contrasts we find that this is due to BIM being worse than the SEE ($t(837) = 4.36$, $p < .001$) and BMS ($t(837) = 2.65$, $p = .063$) models. Taken together, in terms of overall mean predictions we find that all the models perform comparably, and in terms of overall GMD predictions all models except BIM perform comparably. The fact that the SEE model performs comparably to the other models here, while it is the worst model in terms of posterior model probabilities, is an indication that the data is too complex to be captured by simple indices such as the mean or GMD. However, a good model should still be able to capture hallmark features human causal judgments that are described in terms of means or GMDs (Palminteri et al., 2017).

	BMS	PUM	BIM	MVM	SEE
Mean difference	5.481 (0.477)	5.128 (0.344)	6.045 (0.414)	5.344 (0.532)	5.304 (0.420)
GMD difference	5.581 (0.445)	5.880 (0.500)	6.780 (0.555)	5.851 (0.460)	4.804 (0.428)

***Table 5.6** Overall differences between observed and predicted means and GMDs in percentage points. Predictions for each model are obtained by using BMA over the guess components. Standard errors are in brackets.*

Let us consider the qualitative patterns identified in Section 2 next. To investigate the patterns in predicted means we plotted them separately for each inference type and model (Figure 5.5). From the literature we know that people exhibit mean conservatism (pattern 1a in Table 5.2) and Markov violations (pattern 2). Figure 5.5 shows that all models predict mean conservatism, that is, the predictions are between the normative response and 50%. We can see that the BMS most accurately predicts the pattern of Markov violations (first three columns in Figure 5.5), while the BIM appears to predict a Markov violation larger than observed and the other models do not appear to predict Markov violations. Separate ANOVAs for each model on the predictive inferences identify indeed that only for the BMS and BIM there is clear evidence for Markov violations in their predictions (BMS: F(2,56) = 42.7, p < .001, BF_{10} = 68.3; BIM: F(2,56) = 137.2, p < .001, BF_{10} > 1000) with evidence for an absence of Markov violations for the other models (MVM: F(2,56) = 3.32, p = .043, BF_{10} = 0.11; PUM: F(2,56) = 3.02, p = .056, BF_{10} = 0.13; SEE: F(2,56) = 2.94, p = .061, BF_{10} = 0.12).

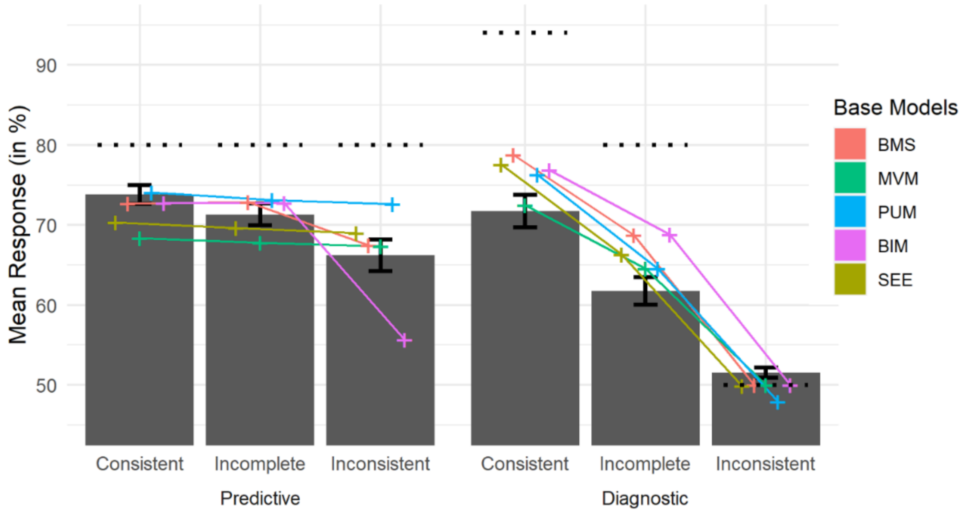


***Figure 5.5** Predicted and observed mean responses per inference type. Bars represent mean responses and error bars their standard error. Crosses indicate mean predictions for each model, these are obtained by using BMA over the Guess Model predictions. Dotted horizontal lines indicate the normative response.*

Next, we look at patterns in predicted variability (indexed by GMD). Specifically, we want to investigate whether the models predict that variability is lower for inferences with incomplete information (pattern 3a) and that variability is higher for diagnostic inferences (pattern 3b). We find that only BIM predicts variability to be lowest for inferences with incomplete information (Figure 5.6), but only the difference with inconsistent information is significant (Δ = 9.34, t(812) = 13.2, p < .001, BF$_{10}$ > 1000), while the difference with consistent information is not (t(812) = 2.05, p = .102 , BF$_{10}$ = 0.998). There is evidence that the SEE model predicts GMD for incomplete information to be lower than for consistent information (Δ = 2.12, t(812) = 2.99, p = .0081, BF$_{10}$ = 7.68), but it is not different from the incomplete information condition (t(812) = 1.03, p = .313, BF$_{10}$ = 0.238). For the BMS, MVM, and PUM models there are no significant differences in predicted GMD due to a change in information.

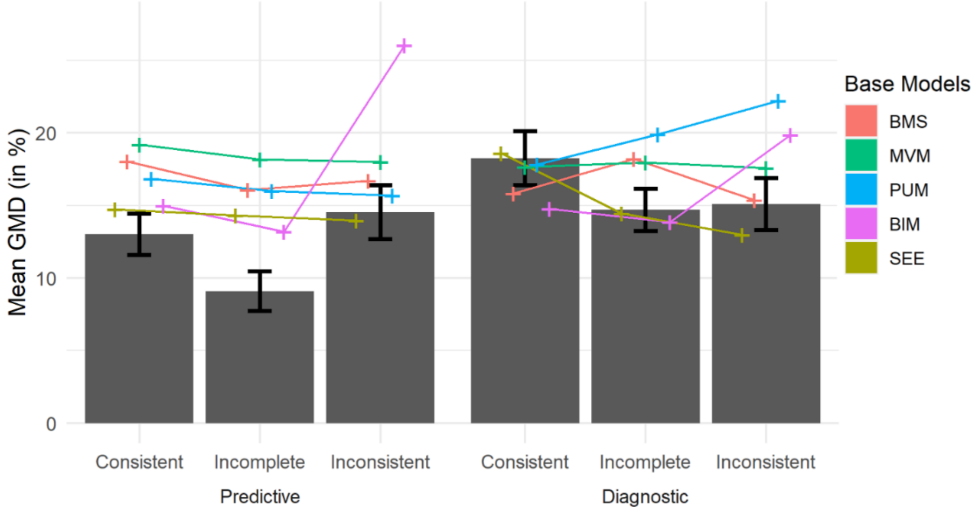


Figure 5.6 *Predicted and observed variability in responses (as indexed by Gini's Mean Difference, GMD) per inference type. Crosses indicate mean GMD predictions for each model, these are obtained by using BMA over the Guess Model predictions. Bars represent the GMD of responses and error bars their standard error.*

Additionally, only PUM predicts that variability is higher for diagnostic inferences (Δ = 3.90, t(812) = 6.73, p < .001, BF$_{10}$ > 1000). No difference between diagnostic and predictive GMD is predicted by BMS (t(812) = -0.89, p = .373, BF$_{10}$ = 0.119), MVM (t(812) = -1.30, p = .193, BF$_{10}$ = 0.267), and SEE (t(812) = 1.72, p = .086, BF$_{10}$ = 0.486), while BIM predicts that the GMD of predictive inferences is higher (t(812) = -3.08, p = .0021, BF$_{10}$ = 9.45).

In all, none of the models adequately captures the observed patterns of within-participant variability over inference types in this dataset. The models fare better at replicating the patterns in mean responses, i.e. mean conservatism and Markov violations. All models predict mean conservatism, but only the BMS and BIM predict Markov violations. Of these two, the BMS clearly fares better at predicting Markov violations and seems so far to be the best account of participant responses. However, not all the qualitative patterns we identified can be observed through indices of central tendency or variability. For these patterns we have to look at the complete response distributions.

5.5.3 Predicted Distributions

To compare the predicted distributions of the models we plotted the observed responses together with model weighted predictions of each base model (Figure 5.7). Note that these predicted distributions are fitted with a single set of parameters, often only 1 or 2, for all 6 inference types. Therefore, an exact fit to the shape of the distribution cannot be expected. However, the successful models should be able to capture the qualitative patterns in responses identified in the introduction.

We find that all the models predict multimodal response distributions (pattern 4). The SEE model predicts three modes for all inferences, but the prediction of a small mode below 50% is not borne out for all inferences. All the other models predict two modes, one near the normative response and one near 50%. The observed distributions mostly follow this pattern, however there are small clusters of responses near 25% for some of the inferences (Figure 5.7A, D, E & F). This is possibly due to participants misreading stimulus values on a few trials (as implemented by SEE). However, these responses seem to be consistently clustered near 25%, not where the SEE model always predicts them to be (this is most clearly visible in Figure 5.7F). This might be due to participants rounding their responses. Such rounding might also cause modes of responses near 75%, something none of the models adequately capture. We return to this observation in the discussion.

All models predict a mode of responses at 50% (pattern 5a). Moreover, all models predict this mode to increase from consistent to inconsistent inference types (pattern 5b) and predict it to be larger for diagnostic inferences (pattern 5c).

BMS and MVM are able to reproduce moderate conservatism, even though they both tend to overestimate responses that are above the normative response (i.e. anti-conservative), as all the models except SEE do. This relates to the lack of extreme responses (pattern 1c). We actually do observe responses near 100%, particularly for the consistent inference types (Figure 5.7A & B). Most models do not predict responses at 100% that are not observed, except the BMS which predicts more responses to be present near 100% especially for the diagnostic consistent and predictive inconsistent inferences (Figure 5.7B & E). While the BMS predicts too many responses near 100% for some of the inferences, the other models, except possibly the PUM, predict too few such responses.

Taken together, the BMS and MVM perform best in terms of predicting these qualitative patterns in the distributions. While neither of these two models capture the distributions perfectly, the other models fare worse. This is in line with the posterior model weights (Figure **5.4**), which indicate that the BMS is the most likely model for the majority of participants, followed by a sizeable group best fit by the MVM. This indicates that the dominant source of variability is different for these two groups of participants, possibly due to using different strategies. A question that arises from this is what differentiates these participants in terms of their behavior. To investigate this, we separately plotted the distributions of responses of these two groups of participants together with their best fitting model predictions (Figure 5.8).

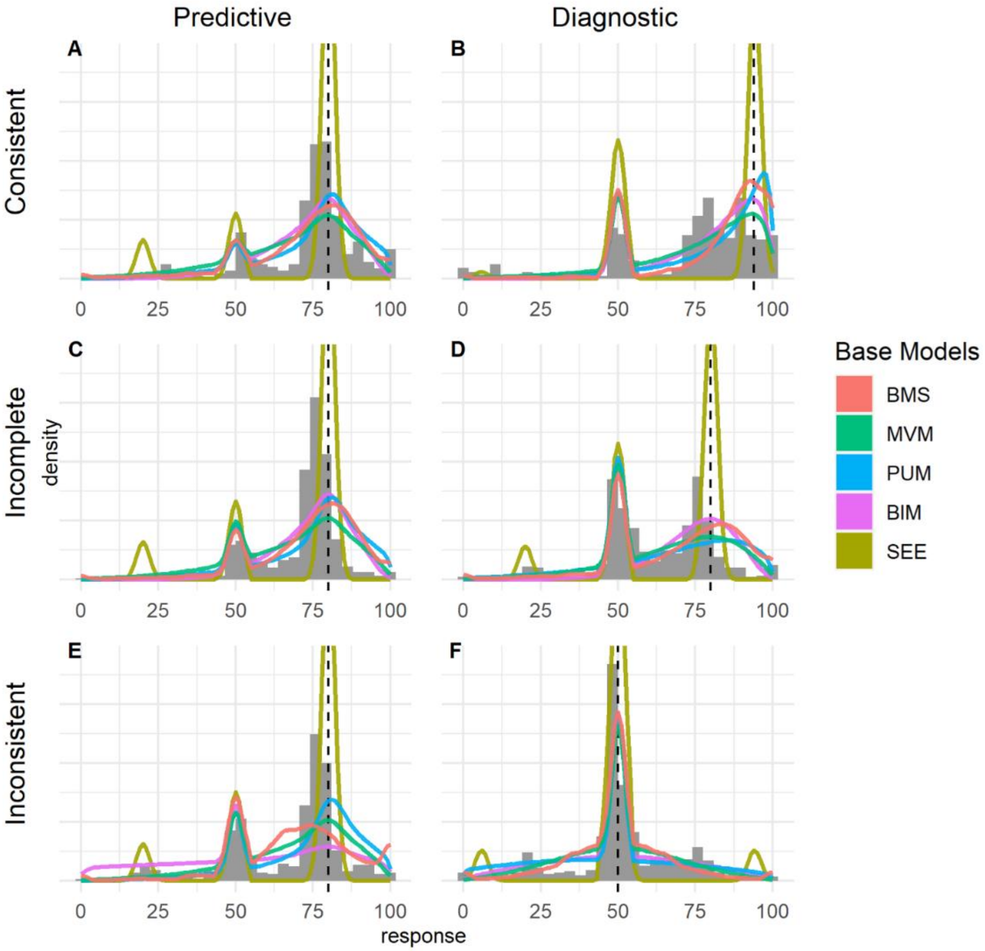


Figure 5.7 *Predicted and observed response distributions per inference type. The grey histograms represent participant responses. The colored lines are the predictions for each base model. These predictions are obtained by model averaging over the Guess Model predictions. Dotted vertical lines indicate the normative response. Because the SEE model predicts substantially more probability mass close to the normative response, the y-axis has been cut-off to allow for better comparison of the densities of the other models.*

Based on the mean responses we have already established that all models predict mean conservatism (pattern 1a; Figure 5.5). Inspecting the predicted distributions for moderate conservatism (pattern 1b; i.e. the bulk of responses fall between the normative response and 50% for all inference types), we find that the SEE and PUM models fail to predict this pattern. This was expected from the SEE model, as it can only predict modes of different sizes at specific values (Table 5.4). For the PUM model, we find it predicts the main mode for predictive inferences to have more mass on the right side (Figure 5.7A, C & E), i.e. it predicts more anti-conservative responses. For the BIM, we find that it mostly predicts moderate conservatism, but not for the predictive inconsistent inference, where it predicts a distribution that is too flat (Figure 5.7E). The

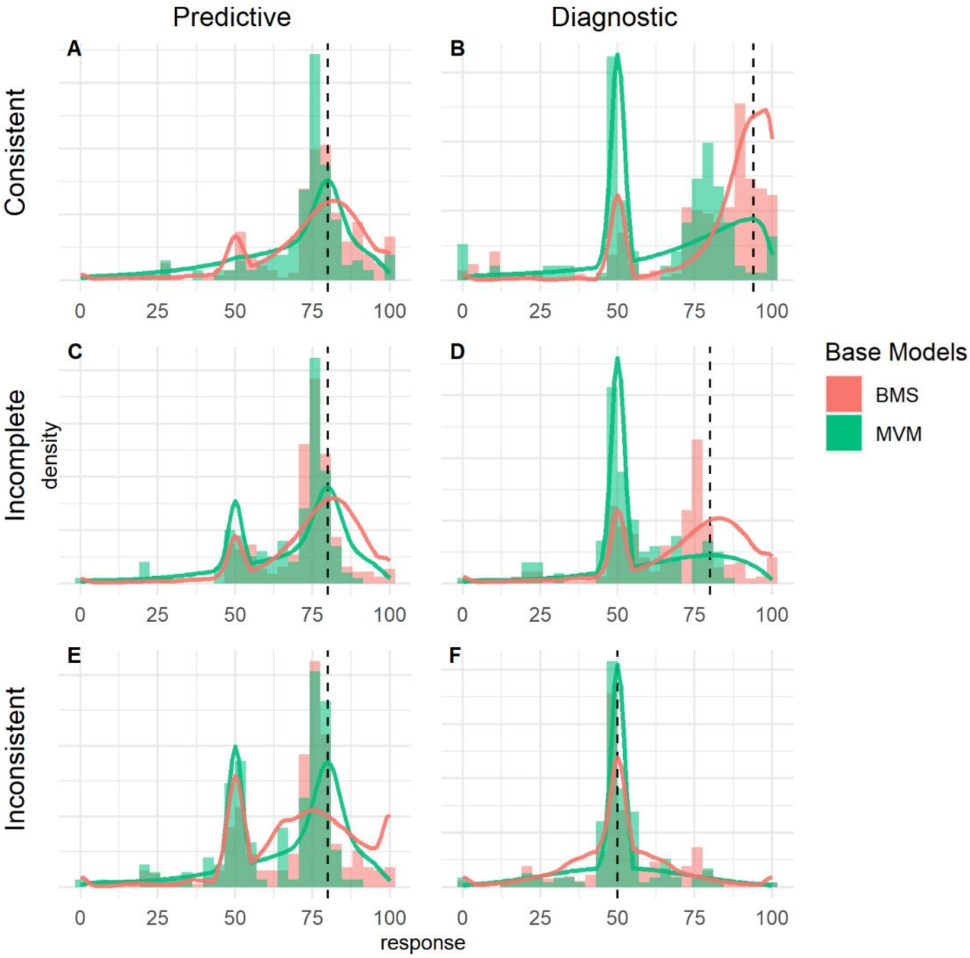


Figure 5.8 *Predicted and observed response distributions per inference type separated for the participants that were best fit by either the BMS or MVM model (25 out of 29 participants). The red and green histograms represent the responses of participants best fit by the BMS and MVM models respectively. The red and green lines represent the model predictions for those groups of participants of the BMS and MVM models respectively. These predictions are obtained by model averaging over the Guess Model predictions. Dotted vertical lines indicate the normative response.*

We observe large differences in the response distributions of these groups for the diagnostic consistent and incomplete inferences (Figure 5.8B & D). For both these inferences we find that the participants best fit by the MVM model are more conservative: the mode at 50% is higher and the second mode is closer to 50%. While the BMS predicts that people are less conservative, neither model accounts for the observed distributions on these two inferences well. We find that both models (and indeed also the other models, see Figure 5.7) overpredict the number of anti-conservative responses. Next, we look at the fitted parameter values, to validate the model fits and further investigate the differences between the groups of participants best fitted by the BMS and MVM.

5.5.4 Fitted parameter values

A summary of the mean fitted model parameters is shown in Table 5.7. For the BMS, we find that the average chain length is within the range expected from the literature, albeit on the higher side (Davis & Rehder, 2020; Kolvoort, Temme, et al., 2023; Zhu et al., 2020). For β we find a similar value as in a previous study (Kolvoort, Temme, et al., 2023), suggesting that participants used a prior distribution of a shape close to the uniform distribution (β = 1). This would entail that participants used, on average, a rather uninformative prior. The SEE model estimates the probability of misreading stimulus values to be on average 11%. While this estimate appears reasonable, it cannot be considered reliable due to the poor model fit. For the PUM, we find that the uncertainty was a lot higher for base rates than it was for the causal strengths. Base rates have a larger influence on diagnostic inferences, as CBN theory stipulates that diagnostic reasoning requires the incorporation of the base rate probability of the cause. So, more uncertainty in base rates leads to more variable diagnostic inferences compared to predictive ones, which is what we observe in participants (pattern 3b) and the PUM predictions (Figure 5.6). This is also in line with previously argued claims that diagnostic inferences tend to be more difficult (Fernbach et al., 2011; Fernbach & Darlow, 2010). We find that the concentrations for the MVM model are higher than for the BIM, which is also expected as the BIM stipulates the concentration to be higher for certain inferences than the fitted parameter (Table 5.3). In all, these parameter values are as expected and indicate the models were implemented and fit correctly.

	BMS		PUM		MVM	BIM	SEE
parameter	Chain length	β	Conc. causal strengths	Conc. base rates	Conc.	Conc.	Error prob.
Mean	70.30	1.218	29.74	347.5	15.07	5.711	0.1081
SD	54.8	1.32	62.6	1542	32.3	12.1	0.0854

Table 5.7 *Estimated mean parameters of base models for all participants. These parameters values are model-averaged using BMA over the Guess Model predictions. ‘Concentration’ is abbreviated as ‘Conc.’.*

We found that participants best fit by MVM were more conservative than participants best fit by the BMS (Figure 5.8). However, as the β prior parameter of the BMS can in principle account for conservatism, it might not be conservatism itself that makes the MVM fit better for those participants than the BMS, but rather that their variability is more characteristic of motor variability. To investigate this, we plotted the best fitting BMS parameters for all participants and color-coded each participant by their winning model (Figure 5.9). The cluster of participants fit best by the MVM model have a higher estimated β parameter (Mean β BMS cluster = 0.91, Mean β MVM cluster = 1.82), but there is little evidence for this difference being significant ($W = 34.5$, $p = .0535$, $BF_{10} = 0.900$). We find no difference between the clusters in terms of their chain length (Mean chain length BMS cluster = 70.1, Mean chain length MVM cluster = 78.8, $W = 73.5$, $p = .770$, $BF_{10} = 0.388$). A higher β would make sense for the MVM cluster as it leads to more conservative responses and this is what we observed for this group (Figure 5.8). As the BMS can predict conservatism with a high β, and indeed we estimate a higher β for this group, it is likely not just conservatism itself that makes that these participants are best fit by the MVM. This would imply that their response distributions are better characterized by motor variability than by sampling variability as in the BMS.

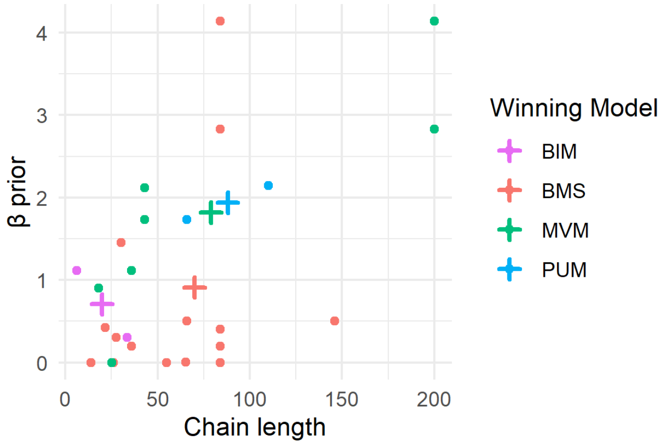


Figure 5.9 *Best fitting BMS parameters for each participant. These parameters are obtained by Bayesian Model-Averaging over the Guess Model predictions. Each dot represents a participant. Dots are colored based on the best fitting base model for that participant. Crosses represent the mean parameter values of the participants in that winning model group.*

5.5.4.1 Guess proportions

Lastly, we conducted an exploratory analysis of the Guess Model and the estimated guess proportions. While the Guess Model was not of primary interest, we provide a short analysis as it can shed light on the necessity of implementing such a mechanism. A guess proportion was estimated as part of the best fitting model for 14 participants, the remaining 15 participants were estimated to not have a guess component (see shading in Figure 5.4). This is evidence for the idea that the observed distributions of responses are due to multiple processes: a reasoning process and guessing. Moreover, multiple participants best fit by the BMS had nonzero guess parameters (Figure **5.4**), implying that the default response mechanism of the BMS cannot by itself account for the large spikes of responses at 50%. To investigate whether the inference types affected the proportion of guesses, we computed model weighted guess proportions for each participant (Figure 5.10).

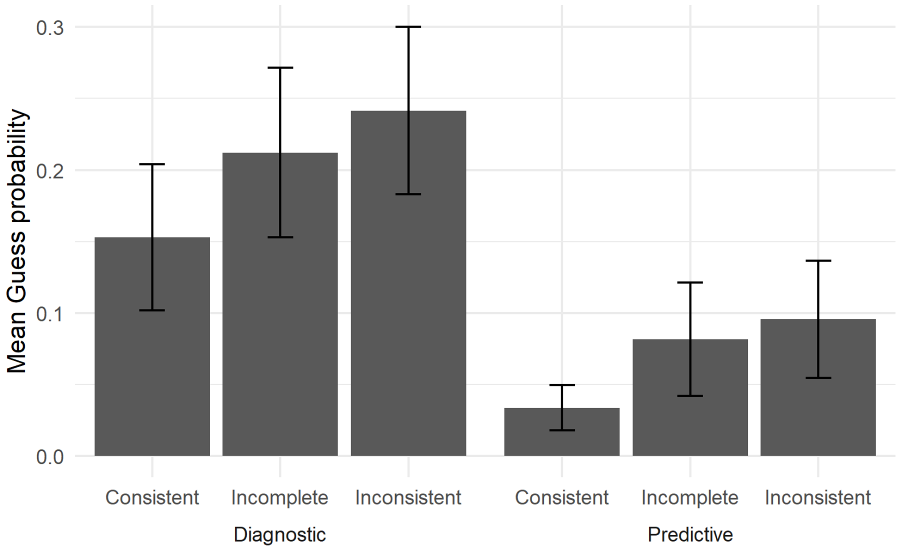


Figure 5.10 *Bayesian Model-Averaged estimates of the probability of guessing for each inference type. Error bars indicate the standard error.*

We find an overall mean guess probability of 0.136 (*SE* = 0.035) which is significantly above zero ($t(28) = 3.94$, $p < .001$, $BF_{10} = 63.1$) indicating that indeed participants guessed on a sizable number of trials. To test whether the direction of reasoning or the provided information affected the amount of guessing we use a repeated measures ANOVA. There is a significant effect of reasoning direction ($F(1, 140) = 22.0$, $p < .001$, $\mathrm{BF}_{10} = 38.9$), with participants guessing more on diagnostic trials (*M* = 0.202, *SE* = 0.053) than on predictive trials (*M* = 0.0703, *SE* = 0.029). This finding squares with previously reported claims that diagnostic inferences are experienced as more difficult (Fernbach et al., 2011; Fernbach & Darlow, 2010). While we observe a trend of more consistent information leading to fewer guesses (Figure 5.10), we find no effect of the information factor ($F(2, 140) = 2.53$, $p = .083$, $\mathrm{BF}_{10} = 0.179$). This latter finding is surprising as we expected participants to guess more when the provided information is incomplete or inconsistent.

5.6 Discussion

Our goal with this study was to disambiguate between theoretical accounts of causal reasoning by leveraging the variability in causal judgments. Out of the set of candidate models we found that the BMS is best able to capture participant responses. In addition, we found that to account for all data it is important to incorporate additional processes, such as guessing. In this section, we will first provide a general evaluation of the candidate models, after which we will discuss the limitations of this work and suggest directions for future research.

5.6.1 Model evaluation

We found that, overall, the BMS has by far the best quantitative fit to the data, being at least 1,000 times more likely to account for all the data than any of the other models. This provides evidence for the claim that the dominant mechanism responsible for the variability we observe is due to sampling from memory or an internal generative model as proposed by the BMS.

Next, on the individual level we found that the BMS and MVM were the best performing models, with 59% of participants best explained by the BMS and 28% by the MVM. Notably, each participant (except one) had a clear winning model. This result indicates that people likely differ in terms of what process is the dominant source of their variability. Moreover, that these quantitative results so strongly favor a single model for each participant indicates that indeed, we can use the variability in causal judgements to disambiguate between theoretical accounts of causal reasoning.

In terms of predicting qualitative patterns in the response data the results were not as clear cut. The models performed comparably and none of the models was able to capture all the qualitative patterns (Table 5.8).

All models failed to account for the within-participant variability patterns over the inference types (patterns 3a and 3b). However, as these patterns require repeated measures data, they are based on a single study and are yet to be replicated, so we may consider these findings preliminary. Nevertheless, a model that would capture these phenomena, would require that certain model parameters are allowed to differ between the inference types. This appears to be a justifiable approach for the BMS, as people may adjust their chain length (i.e. sample more or less) depending on some stimulus characteristics, such as perceived difficulty (Zhu et al., 2020). For example, individuals may choose to generate more samples when faced with a seemingly challenging stimulus in order to improve their accuracy. It is less clear how to psychologically justify varying parameters over inference types for the other models. This would require explanations for why motor variability (MVM), uncertainty regarding underlying causal parameters (PUM), or the probability of misreading part of a stimulus (SEE) would change after observing a stimulus. For the BIM, the ratio of variabilities for each inference are part of the model specification, so it cannot predict diagnostic inferences to be more variable than predictive ones. Therefore, future research that removes the constraint of fixed parameters over inference types appears to be most promising for the BMS.

Qualitative patterns	Predictions				
	BMS	MVM	PUM	BIM	SEE
1a. Mean conservatism	✓	✓	✓	✓	✓
1b. 'moderate' conservatism	✓	✓	X	✓, but not for all	X
1c. Extreme responses are rare (mostly just for consistent inferences)	✓, except for two inferences	✓	✓	✓	✓
2. Markov violations	✓	X	X	✓, but too strong	X
3a. Within-participant variability is lower for incomplete information	X	X	X	✓	X
3b. Within-participant variability is higher for diagnostic inferences	X	X	✓	X	X
4. Multi-modal response distributions	✓	✓, by guesses	✓,by guesses	✓, by guesses	✓, 3 modes
5a. Spikes at 50%	✓	✓, by guesses	✓,by guesses	✓, by guesses	✓
5b. Spikes at 50% increase with inconsistency of information provided	✓	✓, by guesses	✓,by guesses	✓, by guesses	X, not strictly for diagnostic inferences
5c. spikes at 50% are larger for diagnostic inferences	✓	✓, by guesses	✓, by guesses	✓, by guesses	✓

Table 5.8 *Qualitative patterns and model predictions.*

A qualitative pattern that we did not consider from the start, is that the mode of participant responses tends not to fall at the normative probability but is more conservative. We will refer to this as 'modal conservatism'. None of the models predicted this feature of the data. The largest mode for inferences with a normative probability of .8 was near .75, and for the other inferences we also find clusters of responses at .75 (Figure 5.7). Two possible explanations for this are rounding and conservatism. While it is known that people tend to be conservative on tasks with probabilities (e.g. Costello & Watts, 2014; Erev et al., 1994; Hilbert, 2012; Peterson & Beach, 1967; Phillips & Edwards, 1966; Zhu et al., 2020), the fact that we also find clusters of responses at .25 and .5 gives credence to the rounding explanation. That is, participants round their responses to one of three categories, at .25, .5, and .75. Previous work involving probability judgments have made similar observations (e.g. Costello & Watts, 2014; Kleinjans & van Soest, 2014; Wallsten et al., 1993). It seems most likely that both processes are involved. For the diagnostic consistent inference type we find modes just below the normative response and near .75, even though the normative response is .94 (Figure 5.7B), which could be explained by a mixture of rounding at .75 and conservatism. We recommend future research to look at ways for incorporation such a rounding process. Moreover, these observations should serve as a reminder for researchers to take a detailed look at any response distributions. We have previously distinguished ‘mean’ from ‘moderate’ conservatism as empirical phenomena (Kolvoort, Temme, et al., 2023), and it might now be fruitful to add modal conservatism to that list. Being able to predict the location of the mode correctly will likely also alleviate the problem of overpredicting anti-conservative responses that all models suffered from.

One important feature of the data that we wanted to address were Markov violations. The BMS accurately predicted Markov violations, which is a significant advantage of this model over its competitors, given that such violations are a hallmark feature of human causal reasoning (Ali et al., 2011; Davis & Rehder, 2020; Mayrhofer & Waldmann, 2015; Park & Sloman, 2013, 2014; Rehder, 2014, 2018; Rehder & Waldmann, 2017; Rottman & Hastie, 2014, 2016; Sloman & Lagnado, 2015; Waldmann et al., 2008). Although the BIM also predicts the occurrence of Markov violations, it substantially overestimated their magnitude. It was expected that both these models can predict Markov violations, given that they were specifically designed to capture this aspect of human causal reasoning (Davis & Rehder, 2020; Rottman & Hastie, 2016). However, the significant discrepancy between the BIM's predictions and the actual observations suggests that the sampling from the posterior procedure proposed by this model may not fully capture how humans generate causal judgments.

Turning to the spikes of responses at 50%, we discovered that the best fit to the data was achieved using mixtures of the base models with guess components. This finding provides direct evidence that the spikes at 50% are indeed the result of guessing, a hypothesis that had been suggested previously by multiple authors (Kolvoort, Fisher, et al., 2023; Kolvoort, Temme, et al., 2023; Rottman & Hastie, 2016). While the phenomenon of guessing may not be the primary focus of researchers investigating causal cognition, it is nevertheless an important factor that needs to be considered. Our results demonstrate that participants are more likely to guess on certain types of inferences than others, which could introduce bias when analyzing differences in mean responses between inference types. Therefore, it is important to take the effects of guessing or default responding into account when interpreting results and drawing conclusions from causal reasoning experiments.

Taking together both the quantitative fit to the data and the models’ ability to predict qualitative patterns of interest (Palminteri et al., 2017), we find that the BMS outperforms the other models. Our findings together suggest that the process proposed by the BMS is a good candidate for the process by which people generate causal judgements. This would imply that a large part of the variability observed in human causal reasoning data is due to stochastic sampling from memory or a generative model as proposed by the BMS. However, the BMS cannot explain all facets from the empirical data. To explain all the data, it seems necessary to combine different models. In addition to the guess component that we modeled as a mixture here, it might be necessary to combine the BMS with the SEE model or with a rounding process, to explain findings such as the clusters of responses at 25% for multiple inference types. In the next section we discuss more limitations of the current study and suggest additional directions for future research to build upon this work.

5.6.2 Limitations and future research

Even though the BMS seems able to capture human responses better than the other candidate models, it is far from perfect. In addition to the clusters of responses at 25%, the BMS, and all other models, fared badly at predicting patterns of within-participant variability. These patterns, however, were based on data from only a single experiment. This limitation highlights the need for continued empirical and modeling research in this area. Our findings suggest multiple areas in which the study of causal cognition can be improved.

With regard to empirical research, the field needs more studies that involve repeated measures. Our current findings show that it is possible to model the distribution of raw responses (see also Kolvoort, Temme, et al., 2023) and we started with treating within-participant variability as an explanatory target. The repeated measures data we modeled here is currently the only dataset that can be used to investigate such within-participant variability in causal judgments. Further repeated-measures studies could refine our understanding of how the BMS and other models can account for behavior in different experimental settings, and explore patterns of within-participant variability and how they might be better captured. As we argued in the introduction, we should move from group-level mean judgements to participant-level distributions of judgments. Specifically, future studies could validate the patterns of within-participant variability we established previously. In addition, they should look at generalizing and extending these findings. The current study was limited to only six inference types and a single causal structure (a three-variable common cause network) with one set of causal strengths and base rates. Future research could identify new relevant patterns in human judgments and possibly generalize the empirical patterns we identified to other inferences and causal networks. As causal cognition underpins a myriad of judgements in other domains, such as categorization, moral judgements, interventions, and learning (Sloman & Lagnado, 2015), efforts should be made to investigate the generality of behavioral effects in these domains and to investigate generalizing causal models to these domains. The Mutation Sampler, the model of which the BMS is a generalization, has already been shown to extend to categorization and intervention studies (Davis & Rehder, 2020), and we assume this property is inherited by the BMS.

Another type of generalization that is important relates to response scales, that is, the format in which participants are asked to provide a causal judgment. In many studies, including the one we modeled here, this is in the form of a probability judgment. However, the judgements we observe are behavioral reports of some underlying causal belief, but we cannot directly measure that belief. It is still an open question how internal mental representations (i.e. beliefs) map onto a probability (or percentage) scale. Previous studies have argued that indeed people's causal beliefs are graded (Kolvoort et al., 2021; O'Neill et al., 2022), and all studies using a Likert scale, probability, or percentage format implicitly endorse this position. However, it remains unclear how the gradation of causal beliefs relates to gradation on a response scale. It seems unlikely that there would be a one-to-one correspondence. Previous studies have tried to partially avoid this issue by fitting a scaling parameter to map responses to a 100-point probability scale (e.g. Davis & Rehder, 2020) or by changing the response format. Examples of other formats are a frequency formats ('the number of instances out of 20', e.g. Rottman & Hastie, 2016), Likert scales (e.g. T. F. Icard et al., 2017), or having participants choose the most likely causal network state out of two states (e.g. Rehder, 2014). However, none of these options completely avoid the issue of response mapping. One way forward would be to test competing theories on multiple response formats. While labor intensive, such an approach could provide valuable evidence for competing explanations that is to some degree independent of response format.

The aforementioned experimental extensions would allow the candidate theories to be tested on a richer and more varied set of data, allowing for stronger inferences. The current study included variability as a relevant feature of behavioral data, and we focused on full response distributions in a previous study (Kolvoort, Temme, et al., 2023). Other recent studies on causal reasoning have included response times (Kolvoort, Fisher, et al., 2023; Rehder, 2014) and confidence judgments (Kolvoort, Fisher, et al., 2023; O'Neill et al., 2022). We view this as a positive development for the field. The BMS, for example, would allow for the joint modeling of responses and response times, as the chain length parameter is related to response times (Kolvoort, Temme, et al., 2023).

Incorporating more data, however, comes with additional modeling challenges. The simulation-based modeling approach we developed and used here can be extended to incorporate more data sources as well as more (combinations of) models, including ones without a known analytic form. In the current study we were limited by computational resources, which prohibited the inclusion of more parameter values as well as more combinations of models. Future increases in computational resources will help with this and so will new developments in model fitting and evaluation. One promising technique for studying generative models is that of using amortized inference combined with deep neural networks (Fengler & Frank, 2020; Radev et al., 2020, 2022). Amortized inference refers to the process of separating inference and training such that the inference costs are minimized. We applied such a technique here by first constructing a pre-paid grid with all the model predictions (Mestdagh et al., 2019), after which computing maximum likelihoods (i.e. inference) was very fast. Amortized inference allows for re-use, i.e. other researchers can use our grid to fit the candidate models here to datasets from similar experiments. Instead of simulating a grid filled with model predictions, recent approaches train a neural network to learn the mapping between model parameters and predictions. Implementations of this method are now becoming available (e.g. the BayesFlow package for R; Radev et al., 2022) and will allow researchers to pool computational resources to fit a variety of generative models.

5.7 CONCLUSION

Past research into causal reasoning has shown that a variety of computational models can account for different patterns in average causal judgments (e.g. Mistry et al., 2018; Rehder, 2014, 2018; Rottman & Hastie, 2016). There are numerous combinations of these models that could account for all the patterns in average judgements, which puts the field in position from which it is hard to come to a satisfactorily account of causal reasoning as there are too many possible model combinations (Rottman & Hastie, 2016). The current research aimed to make progress on this problem by looking at whether we can use the variability in causal judgment to disambiguate between theoretical accounts of causal reasoning and identify the source of this variability. That is, we considered the variability in causal judgments as something to be explained and not to be averaged out.

Our findings suggest that the sampling procedure proposed by the BMS is a substantial source of variability in probabilistic causal judgments. In addition, our analysis indicates it is important to incorporate ‘non-reasoning’ processes into models of causal reasoning, such as guessing and rounding, to improve the ability to capture human response data.

Overall, this study highlights the potential of computational modeling to illuminate the underlying mechanisms of human causal reasoning, and points to new avenues for experimental and modeling research that could lead to a more comprehensive understanding of this form of cognition.

Part 3

Affordances and Causal Engagement

6 AFFORDANCES FOR SITUATING THE EMBODIED MIND IN SOCIOCULTURAL PRACTICE

Abstract

The Skilled Intentionality Framework (SIF) is a philosophical approach that combines insights from both ecological psychology and enaction to understand the embodied and situated mind. By construing affordances as relations between the sociomaterial environment and abilities available in an ecological niche, SIF radically extends the scope of affordance theory. We propose that it is possible to understand *all* skillful action in terms of engagement with affordances. Moreover, conceiving of affordances in this way allows for an analysis of affordances on multiple scales (e.g. their invitational character for a particular individual as well as the affordances available in a given sociocultural practice) while simultaneously bridging these levels with the SIF to provide an integrated account of the embodied and situated human mind. Our aim in this essay is to showcase these strengths of SIF. In particular, we will discuss the landscape of affordances as our ecological niche; the experience of an individual in a niche structured by affordances; the interrelation of the individual and niche in terms of engagement with affordances; and, lastly, we look at the dynamics within an individual.

This chapter has been adapted from: Kolvoort, I.R. & Rietveld, E. (2022). Affordances for Situating the Embodied Mind in Sociocultural Practice. In Z. Djebbara (Ed.), *Affordances in Everyday Life* (pp. 13-22). Springer, Cham. DOI: 10.1007/978-3-031-08629-8_2

6.1 INTRODUCTION

Much of human daily life is taken up with performing skilled activities in which we engage with the affordances the social, cultural, material, and natural environment provides. Activities as varied as driving, eating, performing surgery, talking, and making works of art can be understood in terms of skilled engagement with affordances. Affordances are possibilities for action provided to us by the environment – by substances, surfaces, objects, and living creatures that surround us (Chemero, 2009; J. Gibson, 1979; Heft, 2001; Stoffregen, 2003). The concept of affordances applies not only to humans, but to all living organisms, as we all share the fate of being inescapably surrounded by our surroundings.

This broad applicability of ecological psychology and its focus on action is shared by enactivism, an approach to cognition that focusses on the dynamic interactions between an acting organism and their environment. The Skilled Intentionality Framework (SIF) is a philosophical approach that combines insights from both ecological psychology and enactivism to understand the embodied and situated mind. With SIF there is the long-term ambition to provide a conceptual framework that applies across the board; to all living organisms, from mollusks to mammals, and to all types of behavior, including so-called 'higher' cognition and collective action. SIF radically extends the scope of affordance theory and in doing so aims to offer a parsimonious account of cognition that provides a sound philosophical foundation for understanding the relation between people and their living environment and, moreover, is relevant for neuroscience, biology, the humanities, and the social sciences alike. The aim of this chapter is to provide an overview of SIF and the role that affordances play in it. Skilled intentionality is the selective engagement with multiple affordances simultaneously, which puts affordances and the responsiveness to them at the heart of SIF.

A cup affords grasping by us, mostly by virtue of physical facts concerning the size and shape of our hands and cups. However, it is possible to explain so much more than just mechanical action routines using affordances if we understand how affordances are related to sociocultural practices. For example, it makes a difference whether a cup is yours or mine: I will be invited by the possibility of drinking from mine but not from yours. Crucially, we propose that is possible to understand *all* skillfull action in terms of engagement with affordances. To accomplish this the SIF proposes a broad definition of affordances as relations between (a) aspects of the sociomaterial environment in flux and (b) abilities available in a 'form of life' (Rietveld & Kiverstein, 2014).

Using this definition allows for an analysis of affordances on multiple scales (e.g. their invitational character for a particular individual as well as the affordances available in a given sociocultural practice) while simultaneously bridging these levels to provide an integrated account of the embodied and situated human mind. Our aim in this essay is to showcase these strengths of SIF and more generally the strengths of a philosophy of affordances that takes our human situatedness in a social, cultural, material, and natural environment seriously. In particular, first we will discuss the landscape of affordances as our ecological niche. Then we discuss the experience of an individual in a niche structured by affordances. In the third part we discuss the interrelation of the individual and niche in terms of affordances. And we end with looking at the dynamics within an individual, namely the bodily states of action readiness that affordances can evoke.

6.2 The landscape of affordances as our ecological niche

The aforementioned definition of affordances uses the Wittgensteinian notion of a '*form of life*' (1953), which refers to "the relatively stable and regular patterns of activity found among individuals taking part in a practice or a custom" (Kiverstein et al., 2019). The reason to use 'form of life' in the definition of affordances is to be able to account for the highly specialized and varied abilities that humans can embody by being part of sociocultural practices. While for most purposes it seems reasonable to characterize the abilities of all members of the earthworm species as a single set, this approach fails for humans, as the skillsets of different individuals, e.g. neurosurgeons and Maasai hunters, vary strongly (see Ingold, 2000). 'Form of life' can thus refer to both sociocultural practices (e.g. those of neurosurgeons or hunters) and to species (e.g. earthworms, kangaroos, humans).

With regard to the environment in which people and other animals are situated, Kiverstein, van Dijk, and Rietveld (2019) proposed to distinguish between the level of the individual and the level of a 'form of life'. At the level of a 'form of life' we can characterize the ecological niche as a *landscape of affordances*. A core idea of the SIF is that the landscape of affordances that surrounds humans is incredibly rich, richer than is generally assumed (Rietveld & Kiverstein, 2014). It is not just that a cup affords grasping; a sad friend affords comforting, this page affords being *described correctly* as white, a surgical room affords a surgeon to do an operation, and a bow and arrow afford the hunter to shoot. Moreover, as affordances are defined relative to a form of life, the existence of affordances is not dependent on the individual. The landscape of affordances is as stable as the patterns of behavior are that form our practices. The landscape thus is a stable, shared environment for individuals inhabiting a form of life (see Figure 6.1A).

The rich human landscape of affordances arises due to the similarly rich relata of our definition of affordances: environmental aspects and abilities available in the form of life. We already touched upon the variety in human abilities; the wide variety of human sociocultural practices entails many different abilities that can be available to human individuals. The other relatum, the environmental aspects, come in even greater variety and are in the human case best understood as being thoroughly *sociomaterial* due to the intertwinement of the material and the social in practice (van Dijk & Rietveld, 2017). As humans we are embedded in sociocultural practices which means that also the material structures around us have been shaped by cultural practices. Wherever you are now, look around and you will see particular objects in particular places, both those objects themselves and the places they are in have been formed by social practices (e.g., this shows itself in that we tend not to put mugs on top of keyboards or keyboards on chairs).

As both our abilities and our environments come about through sociocultural practices, it follows that the landscape of affordances for humans is also fundamentally *social*. The possibilities for action we have depend on the sociocultural practices, i.e. forms of life, we are part of. For example, as part of the sociocultural practice of speaking English, we have to possibility to judge the arguments in this text, to imagine how it could be structured differently, to read out these words aloud, etc. The landscape of affordances in this way reflects the abilities that arise from our practices.

These abilities that arise from our practices include those which have been related to so-called ‘higher cognition’, such as judging the arguments in this text. While research in embodied cognition has mostly focused on sensorimotor skills, we contend that responsiveness to affordances is not limited to repeating mechanically some routine, but is flexible in a context-sensitive way. The orthodox dichotomy of so-called ‘higher’ and ‘lower’ cognition hence plays no role in the SIF; all skilled behavior is viewed as engaging with multiple affordances, enabling the analysis of all forms of behavior in one framework. This includes activities such as reflecting, judging, imagining, verbalizing, planning and more (Kiverstein & Rietveld, 2018; Kolvoort, Schulz, et al., 2023; Van Den Herik & Rietveld, 2021; van Dijk & Rietveld, 2021a, 2021b).

We can think of ‘higher’ cognition as part of temporally extended activities in which we coordinate with nested affordances in an environment structured by a complex constellation of sociomaterial practices (Kiverstein & Rietveld, 2018; van Dijk & Rietveld, 2021a).

Crucially, using the form of life as the level of analysis allowed the development of a Wittgensteinian notion of *situated normativity* to describe the normative aspect of cognition in skillful action (Rietveld, 2008). Situated normativity describes the normative dimension of the things we do in real-life contexts. In every concrete situation an individual distinguishes between better or worse possibilities for action. For humans this is strongly dependent on the sociocultural practices in which our actions are embedded, whether some action is adequate (or good, correct, etc.) or not, is dependent in part upon agreement in action among members of a sociocultural practice (Wittgenstein, 1953). While dancing might be laudable within the confines of a nightclub, it might not be so when engaging in the practice of listening to a client’s presentation at a company office.

6.3 INDIVIDUAL EXPERIENCE OF AFFORDANCES

We have discussed that we can describe the ecological niche as a landscape of affordances on the level of a form of life. An important question is how an individual engages with this landscape. As the landscape of affordance is relative to a whole form of life, this question narrows to: How does an individual *selectively engage* with affordances that are relevant to them in their current situation? If we walk into a cafeteria looking for a place to sit and eat our lunch, we tend not to be overwhelmed by the myriad of possibilities that the chairs, tables, and people in the cafeteria afford us. In such a situation, we tend to be drawn in, or *solicited*, only by aspects of the cafeteria that will allow us to sit down and eat.

In SIF *solicitations* are distinguished from affordances (Rietveld, 2008; Rietveld & Kiverstein, 2014), where solicitations are those affordances that are experienced as *relevant* by a situated individual. So, these solicitations or *relevant affordances* are to be analyzed at the level of the individual, while available affordances and their existence belong at the level of a form of life.

What makes one affordance relevant but not another? SIF argues for a *process of self-organization* as the source of relevance (Bruineberg & Rietveld, 2014). All organisms tend towards a state of relative equilibrium in the dynamic coupling between their body and the world via "self-organized compensatory activity" (Merleau-Ponty, 2003). It is this tendency that imbues some affordances with relevancy but not others and the SIF characterizes this tendency as a *tendency towards better grip* on the situation. It is those affordances that allow us to improve our grip on the situation that are relevant. Which is why in the previous example we are solicited by what an empty chair affords in a cafeteria, but not by the affordances of chairs with occupants.

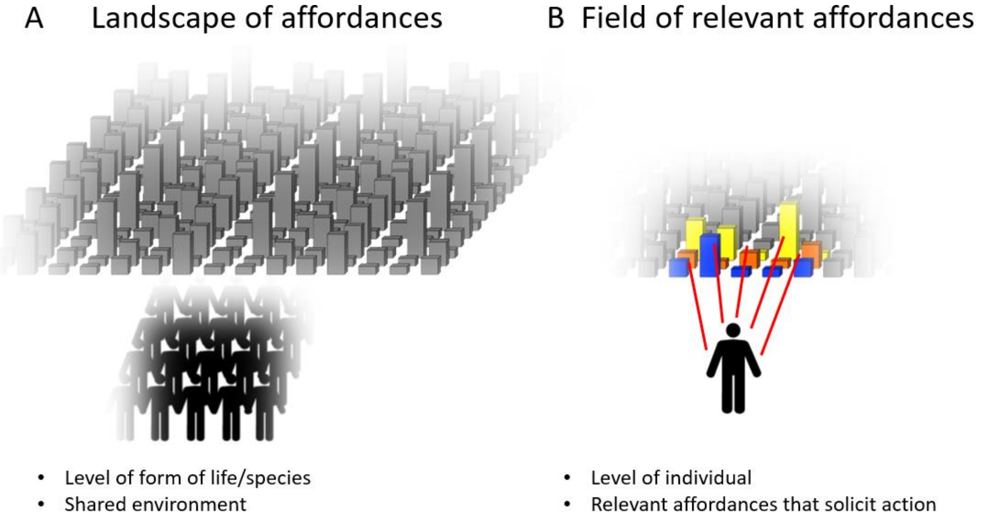

A Landscape of affordances
B Field of relevant affordances
• Level of form of life/species
• Shared environment
• Level of individual
• Relevant affordances that solicit action

Figure 6.1 *Sketches of landscape and field of affordances, which are relative to a form of life and to an individual respectively. Note that the landscape and field are both dynamic (see main text). The field and landscape stand in mutual and reciprocal dependence to one another (Kiverstein et al., 2019).*

However, in real-life we do not engage with only one affordance at a time, Skilled Intentionality implies a responsiveness to multiple affordances simultaneously. We refer to the constellation of affordances that are relevant or inviting to an individual engaging with a concrete situation as the *field of relevant affordances* (Figure 6.1B; Rietveld et al., 2018; Rietveld & Kiverstein, 2014). The inviting affordances of the field are part of the lived experience of an individual (Withagen et al., 2012), and it is opened up out of the landscape, by their abilities and concerns in the concrete situation. This experience of a situation inviting behavior goes together with a *bodily state* that has been referred to as "action readiness" in emotion psychology (Frijda, 2007), that is, the body poises itself for active engagement with relevant affordances.

The landscape of affordances is in flux when considered over larger timescales, but the field of relevant affordances is an even more dynamic and ever-changing phenomenon. When an individual acts or when the situation itself develops, the individual-environment relation is changed and other solicitations arise (Bruineberg & Rietveld, 2014). What is foreground and what is background shifts continuously, the field is in *flux* over shorter timescales. Crucially, the individual is responsive to field of relevant affordances *as a whole*. For example, while attending a presentation, we can be responsive to what is afforded by our cup coffee and the speaker at the same time. And the relevance of what either affords can change due to our own actions (e.g. finishing the coffee, raising our hand) or by the changing environment (a colleague walking in, the presentation ending). Being poised for multiple relevant affordances simultaneously allows for an improvement in grip, because it enables one to flexibly and rapidly respond to changes in the environment (Bruineberg et al., 2021).

6.4 THE INDIVIDUAL ENTANGLED WITH THE FORM OF LIFE: FIELDS AND LANDSCAPE AS CONTINUING PROCESS

Now that we have discussed the landscape and field of affordances, we can turn our eye to their complex and dynamic interrelationship. While we can conceptually distinguish shared publicly available affordances and those relevant affordances that invite a situated agent to act, they should not be separated on ontological grounds (Kiverstein et al., 2019). Such an ontological separation would violate the reciprocal and mutual dependence of the landscape and field. This violation becomes clear when we appreciate the fact that while the landscape of affordances incorporates physical and material structure, it is not the reality as described in physics. Instead, the landscape of affordances is pragmatically structured by patterns of regular activity available in an ecological niche or form of life.

For example, while it is indeed a physical matter that we are supported by the floor of a post office, that we often form a single file queue is not just a physical matter (as the physical space would allow a group to stand in a myriad of configurations), but it is a matter of sociocultural practices, in this case the practice of queuing. Queueing is a practice, it is a pattern of regular activity available in a form of life (one that most of us inhabit), hence it is part of the sociomaterial landscape of affordances. From the perspective of the individual, queueing is also an act, it is an individual engaging with a relevant affordance. This points us towards the reality that practices and affordances are different perspectives on the same thing. The practice of queueing consists out of individuals who tend towards better grip on their situations by engaging with the affordance to queue. When we take the perspective of one individual who enters the post office, the other individuals queueing form part of the sociomaterial structure around her, constraining her field of relevant affordances. When she joins the queue, she engages with the practice of queueing available to all the people there as part of the landscape of affordances.

We chose the example of queueing because of its physicality, as one person queueing (engaging with a relevant affordance) in a very physical sense is both part of a practice (landscape) and a relevant affordance for another person (field), who can queue physically behind her. In a very direct sense the material structure of the landscape (a queue) is here entangled with patterns of an individual's activity. However, this mutuality of practices and affordances is not restricted to physical (material or temporal) contiguity. For instance, the contours of streets have been shaped by practices of people traveling in different ways (e.g. by foot or car) and by builders placing things in certain places (e.g. traffic lights, sidewalks, buildings), which determine the structure of the landscape for everyone who travels that street, even decades later.

From these examples we can learn that practices and affordances are perspectives on the same sociomaterial entanglement of people, actions, places, and things. Activities are related to practices in a fundamental sense (van Dijk & Rietveld, 2017). The practice of queueing exists by virtue of individual acts of queueing. The landscape of affordances is formed partly by a history of individual (or joint) activities and continues to take shape as practices unfold. On the other side of the coin, we have that individual acts of queueing depend on the existence of the practice of queueing. The field of relevant affordances opens up out of the landscape.

This reciprocal dependence between the landscape and field of affordances necessitates a view in which an ongoing process shapes the landscape and field together (Kiverstein et al., 2019; van Dijk & Rietveld, 2021a). This ongoing process is comprised of the activity of individuals: Individuals, enacting relevant affordances, simultaneously shape their field of relevant affordances as well as contribute to sociomaterial practices that shape the landscape of affordances (which in turn will shape the future history of activity of individuals). This process view points towards a temporal view on the relation between the landscape and field of affordances (Kiverstein et al., 2019). On short time scales, the more stable landscape constrains the affordances available in the more dynamic field. For instance, the affordance to queue when one gets to the post office is made possible by existence of the practice of queueing, which exists on a larger temporal scale than a particular individual engaging with the affordance to queue. Over longer periods of time, however, the landscape depends on the field of relevant affordances. Practices are maintained over time by the inviting character of affordances leading to activities constitutive of the practice. The practice of queueing is maintained by virtue of the soliciting character of the affordance to queue to individuals. Individuals queueing keep the practice of queueing “alive”. In this way the field, which invites individuals to act in concrete situations, is “at the forefront” of the evolving landscape, continuing it through time, maintaining it how it is or evolving it in new directions (Kiverstein et al., 2019; van Dijk & Rietveld, 2021a). Kiverstein, Van Dijk, and Rietveld offer the example of musicians making jazz: “the affordances of musical instruments to make jazz music depends upon musicians that know the history of jazz, and can maintain this history whilst also building on it through their own improvisations.” (2019, pg. 2293).

It is important to note that some of the real-world examples we discussed above (e.g. queueing) can perhaps be considered somewhat trivial. These examples were chosen to be familiar and accessible, but considering our claim that all skillful activities can be understood in terms of engaging with affordances, one can expect SIF to be able to do more. One (not so familiar) example of applying the SIF is the analysis of the field of relevant affordances of patients receiving deep brain stimulation (De Haan, Rietveld, Stokhof & Denys 2013). More generally, to understand complex and temporally-extended engagements in terms of affordances requires the methods of embedded philosophy and longer-term ethnographic observation. Examples of using these methods combined with SIF include the practices of psychiatry (van Westen et al., 2019, 2021), visual art and architecture (Rietveld & Brouwers, 2017; van Dijk & Rietveld, 2021a).

6.5 WITHIN THE INDIVIDUAL

So far we have regarded an individual's actions and the dynamics of a developing situation as impacting the individual-environment relation, but the SIF also connects these phenomena with the ongoing dynamics *within* an individual's body and brain. Employing principles from the complex and dynamical systems literature, the SIF relates phenomenology and ecological psychology to developments in theoretical neurobiology (see Bruineberg, Kiverstein, et al., 2018; Bruineberg & Rietveld, 2014, 2019b).

The improvement of grip on a situation can be characterized as the reduction of disequilibrium in the 'brain-body-landscape of affordances' dynamical system. Organisms selectively engage with those affordances that reduce its disequilibrium with the environment. The SIF views this disequilibrium as a *dis-attunement* between internal and external dynamics, i.e. between self-organizing affordance-related states of action-readiness in the individual and the changing landscape of affordances (Bruineberg & Rietveld, 2014). It is this dis-attunement that as a most basic concern drives organisms to selectively engage with relevant affordances. On SIF's view, Friston's Free Energy Principle (2010) is all about improving grip on the field of affordances, a reduction in free energy is a reduction in dis-attunement of internal and external dynamics (Bruineberg, Kiverstein, et al., 2018; Bruineberg & Rietveld, 2014).

Importantly, this conceptual scheme allows for cross-fertilization between disciplines: the study of activity in the brain and body can inform and be informed by investigations of an individual's landscape of affordances (including the embedding sociomaterial practices, which can be investigated well by means of ethnography, see van Dijk & Rietveld, 2021a) and the structure of the field of relevant affordances (which incorporates the individual's abilities). Overall, we contend that to understand the situated mind, we need to regard the whole *brain-body-landscape of affordances* system.

7 AN AFFORDANCE-BASED ACCOUNT OF CAUSAL ENGAGEMENT

Abstract

Causal cognition is a core aspect of how we deal with the world, however existing psychological theories tend not to target intuitive causal engagement that is done in daily life. To fill this gap, we propose an Ecological-Enactive (E-E) affordance-based account of situated causal engagement, i.e. causal judgments and perceptions. We develop this account to improve our understanding of this way of dealing with the world, which includes making progress on the causal selection problem, and to extend the scope of embodied cognitive science to causal cognition. We characterize identifying causes as selectively attending to the relevant ecological information to engage with relevant affordances, where these affordances are dependent on individual abilities and context. Based on this we construe causal engagement as based on a learned skill. Moreover, we argue that to understand judgments of causation as we make them in our daily lives, we need to see them as situated in sociocultural practices. Practices are about doing, and so this view helps us understand why people make these judgments so ubiquitously: to get things done, to provide an effective path to intervening in the world. Ultimately this view on causal engagement allows us to account for individual differences in causal perceptions, judgments, and selections by appealing to differences in learned skills and sociocultural practices.

This chapter has been adapted from: Kolvoort, I.R., Schulz & Rietveld, E. (In press). The Causal Mind: An Affordance-Based Account of Causal Engagement. *Adaptive Behavior*.

7.1 INTRODUCTION

One fundamental way in which we humans experience and deal with the world is by way of causal relationships. This seems to be true in any situation. Whether we are confronted with a scenario involving billiard balls colliding or a social setting in which a friend responds emotionally to someone else's remarks. When we encounter worldly events we perceive more structure than meets the eye (or any other sensory organ). To us it is not just that one billiard ball starts rolling after the other stops, it is not just that our friend becomes emotional after another's words. Instead, it seems central to the way we cope with the world, both individually and as communities, that we experience that one ball *caused* the other to move and that someone's words *caused* an emotional reaction.

This is the phenomenon under consideration here, that of an individual perceiving, judging, and selecting causes of concrete encountered happenings in the world. In the literature these phenomena tend to be referred to as causal perceptions or causal judgments, however our account targets something more basic that encompasses both perceptions and judgments. We focus on the type of causal cognition that is intuitive and forms in the relation between agent's environment and her actions, a type of causal cognition that is ubiquitous. We will use the term 'causal engagement' for this. This paper has three related aims. The main aim is to develop our understanding of the psychology of causal engagement, and the sub-goals are to make progress on the causal selection problem and to extend the scope of embodied cognitive science.

To improve our understanding of causal engagement we will provide a philosophical analysis of the psychological processes that underlie this way of dealing with the world and elucidate why we perceive some things to be causes but not others. Understanding this aspect of our lives, we will argue, requires an affordance-based account, where affordances are the possibilities for action provided to us by the environment (Chemero, 2009; J. Gibson, 1979; Kolvoort & Rietveld, 2022; Rietveld & Kiverstein, 2014).

In providing such an affordance-based account we extend the scope of embodied cognitive science to a core facet of so-called "higher" cognition. Our account is part of the larger literature using the framework of embodied and situated cognition. Embodied and situated approaches to cognition are starting to be applied to more and more facets of cognition. Initially these accounts focused on what has been called "lower" cognition, such as perception or mechanical action routines. More recently, however, much work has been done to extend the scope of embodied and situated accounts to so-called "higher" cognition. Embodied accounts have made headway in understanding imagination (Gallagher, 2017; van Dijk & Rietveld, 2020), mathematical cognition (e.g. Abrahamson et al., 2020; Zahidi & Myin, 2016), anticipation (e.g. Jurgens & Kirchhoff, 2019; Stepp & Turvey, 2015; van Dijk & Rietveld, 2021a), change-ability (Rietveld, 2022), language (Atkinson, 2010; Kiverstein & Rietveld, 2021; Van Den Herik, 2018; van Dijk & Rietveld, 2021b), and more. These works generate doubt about the veracity and productivity of the higher-lower cognition dichotomy and help make sense of the mind using a unified approach. We continue this trend here by providing an embodied and situated account of a core component of "higher" cognition.

This paper is organized as follows. In Section 7.2 we will introduce the causal selection problem and existing perspectives on causal cognition, both of these will illustrate the need for an embodied and situated account of how we engage with causality. Next, in Section 7.3 we will introduce concepts from the Ecological-Enactive (E-E) framework that we will use to build our account. In Section 7.4 we will introduce interventionism as a natural starting point of an embodied account of causal cognition. Then, in sections 7.5 to 7.7 we construct our account of causal engagement in three parts: Section 7.5 focuses on how agents *identify* parts of the environment as causal. In Section 7.6 we discuss what *causality* and *causal relationships* are from the perspective of an agent. Lastly, in Section 7.7 we analyze *interventions*, i.e. the actions we take that are based on and impact the causal systems around us. We conclude the paper with a short summary and we suggest directions for future research based upon the theory developed herein.

7.2 Causality in philosophy, psychology, and life

To set the stage before developing our own account it is important to have a preliminary discussion of some of the relevant literature on causality. To restrict the scope of our account we first discuss the distinction of 'actual' and 'general' causation. Next, we introduce the causal selection problem and discuss an important account of it that indicates how we can me make progress on it. Lastly, we discuss prominent theories of causal cognition in the psychological literature and empirical findings that point towards the need for further theoretical development.

7.2.1 'Actual' causation encountered in the environment

The literature on causality commonly distinguishes two forms: actual and general causality. *Actual causation*[22] is about concrete cases. Judgments of actual causation come about by asking "What is the cause of this?", where 'this' refers to an actual, concrete event that happened in the world. An example of this is "Did Jane's fatigue cause the traffic accident?". This can be contrasted with *general causation*, which is about which causal relationships hold across multiple instances, e.g.: "Does fatigue cause traffic accidents?".

[22] Other names used for this phenomenon are token or singular causation (see Danks, 2017).

As we are mainly interested in cognition situated in daily life, our analysis will be mostly restricted to actual causation. These causal judgments occur when we care about the causes of a specific event and tend to be more intuitive than judgments that require generalization. In daily life we often care about causes of particular events in our environment. This makes judgments or perceptions of actual causes ubiquitous in everyday life (‘What caused Mark to decline my invitation?’), but also in more formal settings, such as medicine (‘What is the cause of this inflammation?’), legal settings (‘What is the cause of the criminal’s actions?’), engineering (‘What caused this bridge to collapse?’), and many others.

7.2.2 The causal selection problem

Understanding how people perceive and judge causes is closely related to the problem of causal selection. The problem of causal selection has received attention from philosophers for many decades and concerns what we *should* pick out as ‘the cause(s)’ of an event out of the many possible causes (Hesslow, 1988; Lewis, 1974). Logically speaking any event has infinitely many causes. We can, for example, trace back a causal chain as far back as the big bang for any event. This has led multiple philosophers to view causal selection as objectively groundless (e.g. Lewis, 1974), but the philosophical work on the problem is still helpful in informing our descriptive account.

A famous example discussed by Carnap (1966, pp. 191–192) illustrates an important feature of causal selection in real life, namely that it can vary strongly:

> EX1: An angry driver is speeding down a street while it is raining. While turning a corner he hits a bump, the car spins and crashes into a wall. What was the cause of this car crash? Carnap claimed that we should not expect a consensus regarding the cause of the crash as different people will focus on different aspects. A policeman might attribute the crash to the driver speeding, while an engineer would point to the state of the road, and a psychologist would focus on the driver's mental state.

So it seems that there are an infinite amount of causes to select, and people tend to select different causes. These facts seem pertinent to any theory of causal cognition. While much progress has been made in understanding causal selection, it is still unclear how and why people make different causal selections.

Hesslow (1988) has argued famously that we should see differences in these causal attributions as differences in questions asked, i.e. differences in the object of comparison. For example, the question “What caused this house to burn down?” could refer to “What caused this house, but not the one next door, to burn down?”, but it could also refer to “What caused this house to burn down now and not yesterday?”. These questions are different, they involve different comparisons. Pointing out a cause that involves the building materials of the house is appropriate for the former question but not for the latter as they probably did not change from yesterday to today. Hesslow (1998) thus proposes that people select different causes because they are actually asking different questions. Unfortunately, no proper explanation is provided of what makes people ask these different questions. Why did the policeman and engineer 'ask different questions' and thus select different causes? Hesslow puts it down to what he calls 'subjective' and 'unconscious' factors such as experience, norms, and education, but provides no account as to how those factors lead to differing causal judgments. This is unfortunate as getting that process in view would help us understand what causes people select and why they do so. We aim to fill this gap with our account by providing more guidance on how and why factors such as education, learned abilities, and sociocultural practices affect causal selection.

7.2.3 The psychology of causal cognition and attribution

While philosophers have debated what makes a cause a good cause to be selected, psychological theories have focused on what information people use and how they use it to make causal judgments. The most prominent theories come in two flavors, they either focus on how individuals learn and reason from statistical dependencies (also called difference-making theories, e.g. Causal Model Theory; Sloman, 2005) or from considerations of (physical) forces (e.g. Force Dynamics; Wolff, 2007). Roughly speaking, the former posits that A causes B if the occurrence of A increases the probability of the occurrence of B, while the latter holds that A causes B if A transfers some physical force to B. It is certainly true that statistical and force considerations affect causal judgments and it has been argued that both are involved in our causal cognition (e.g. Glymour et al., 2010; Lombrozo, 2010; Waldmann & Mayrhofer, 2016). However, it is also clear that these two criteria do not provide the full story. Appealing to considerations of statistical dependency or of forces will not help us pick a cause in the car crash example (EX1) nor provide guidance on why the different agents pick different causes. There are too many possible causes that fit the criteria of dependence and transference. For instance, there is both a dependency and force relationship between the crash and the invention of the combustion engine, but this invention as such is unlikely to be picked out as the cause. Hence applying these criteria would give us a list of candidate causes that is too long to be useful, which means that these accounts suffer from too much underdetermination in concrete situations.

Reducing causality to a single objective criterion, whether it be statistical co-occurrence or transference of force, necessarily leads one to abstract away from experience and the context in which causality is judged (Bender, 2020). While such isolation is essential to science, it can hamper appreciating more complex phenomena. Instead of isolating the psychological phenomena of causality purely in terms of cognition, information, or logic (see Dutilh-Novaes, 2019), we need to regard the full human-environment system in order to more fully appreciate how causes play a role for the human mind.

Empirical evidence points us this way too. Multiple experiments have shown that context (like culture) is incorporated into causal judgments (Bender et al., 2017; Bender, 2020; I. Choi et al., 1999; McGill, 1995; Morris et al., 1995) and developmental evidence indicates the interconnectedness of causal cognition and concrete motor abilities (Sommerville et al., 2005). These facts seem pertinent to any theoretical account that tries to elucidate how people make sense of and use causes in daily life.

The fact that the aforementioned theories are decontextualized reflects their narrow scope: While most authors state the fact that causal reasoning is ubiquitous in human life, the experimental methods used in this field mostly require participants to think reflectively about abstract causal relationships, of (possibly) abstract events or variables, in an abstract laboratory setting. This is in stark contrast with the intuitive manner in which we deal with causal relationships in our daily lives. When someone asks you “What caused you to be late?” or “What caused John to be sad?”, do you really always reflect on the set of possible causes? We think not. This is not to say that people do not have this ability. Experimental evidence clearly shows that they do, it is just that often such reflection is not at play. This divide between reflective and intuitive causal reasoning is also suggested by developmental data indicating their separate development (Kuhn, 1989; Muentener & Bonawitz, 2017a). However, we will not attempt to provide or promote a clear separation of these processes. Instead, we focus on understanding the phenomena of engaging with actual causes in daily life, which is often more an intuitive than a reflective phenomenon.

Before developing our account, it will help review some of the core principles of the E-E framework as these principles are the foundation on which we build our account in later sections.

7.3 THE ECOLOGICAL-ENACTIVE FRAMEWORK

Our account will combine insights from the fields of ecological psychology and enactive cognition (see Kolvoort & Rietveld, 2022). The central notion behind the *enactive approach to cognition* is that perception is something an organism *does* (Froese & Di Paolo, 2011; Gallagher, 2017; McGann et al., 2013; Myin, 2016; Noë, 2004, 2012). In this tradition cognition has been defined as ‘perceptually guided action’(Varela et al., 1991) with action and perception part of the same ‘perception-action loop’ (Stewart, 2010).

Ecological psychology also appreciates the inherent relationship between action and perception. The core concepts underlying this ecological approach are *affordances* and *ecological information* (J. Gibson, 1979). Affordances refer to action possibilities provided to an organism by its environment and they are central to the ecological view on perception: organisms do not perceive the world in a way separated from themselves, instead they perceive the action possibilities the environment affords them. Which affordances are perceived is dependent upon aspects of both the organism and environment. The abilities or skills an organism has are crucial here, as it is those abilities that allow it to interact with the environment in a specific way. Hence affordances are relative to what an organism can do, they are relative[23] to their abilities (Heft, 1989; Kolvoort & Rietveld, 2022; Rietveld & Kiverstein, 2014). This view of affordances allows for expanding the explanatory scope of affordances to include all skillful behavior[24] (Bruineberg, Chemero, et al., 2018; Kiverstein & Rietveld, 2018, 2021; Rietveld et al., 2018; Rietveld & Kiverstein, 2014; van Dijk & Rietveld, 2021b).

Ecological information refers to the regularities and structures present in the environment that enable an organism to engage with affordances (J. Gibson, 1979). To expand the traditional scope of ecological psychology Bruineberg, Chemero, and Rietveld introduced the notion of *general ecological information* (Bruineberg, Chemero, et al., 2018), which refers to the structures and regularities in the *sociomaterial* environment. By encompassing material aspects of the environment, this notion takes into account law-like regularities we find due to our world being governed by physical laws. Crucially though, general ecological information also encompasses the social environment, and thus takes into account regularities that an individual encounters due to sociocultural practices. We will see later that these practices are an important component in understanding judgments and perceptions of causation.

7.4 INTERVENTIONISM: THE NATURAL STARTING POINT FOR AN ECOLOGICAL-ENACTIVE ACCOUNT

Using the empirical facts and concepts discussed in previous sections, we can now start building our affordance-based account of causal engagement by discussing the interventionist theory of causality.

[23] There is a long debate over whether affordances are best treated as relations between organism and environment, as we do, or as dispositional properties of the environment. For the latter view see (Scarantino, 2003; Turvey, 1992).

[24] As mentioned in the introduction, traditionally affordance-based analyses focused on so-called “lower” cognition, such as the perceptuomotor routine of grasping a glass or climbing stairs (for a seminal example see Warren, 1984). Recent work has argued for a much broader conception of affordances (Kiverstein & Rietveld, 2018, 2021; Rietveld & Kiverstein, 2014) that can be used to understand all skillful action, which is in line with the observation by Gibson (1979) that affordances comprise “the whole realm of social significance” (p. 128) in the human form of life. In this paper we build upon these conceptual developments. However, it is important to note that there is no consensus on the scope of the concept of affordances and this topic is still highly debated (for alternative views we refer the reader to Golonka, 2015; Golonka & Wilson, 2019; Turvey, 1992; Turvey et al., 1981).

Philosophers have developed various interventionist[25] accounts of causation (Hitchcock, 2012; Hitchcock & Knobe, 2009; Menzies & Price, 1993; Pearl, 2009; Woodward, 2005, 2014, 2016) which share the same core principle: causes are like handles in the world, that can be acted upon and used to manipulate the world. It is because of this core principle that interventionism is a natural starting point for an ecological and enactive perspective on causal cognition, it puts action immediately on the center stage.

Interventionism was developed as a philosophical account of what causation is. It posits that what it means for 'X to cause Y' is that 'bringing about X would be an effective means to bring about Y' (Menzies & Price, 1993). Otherwise put: X causes Y if and only if intervening on X changes Y.

While many critiques of interventionist theories of causality have been offered (see Price, 2017; Woodward, 2016), these are not inherited by our proposal as we are not offering an account of the epistemology or metaphysics of causation itself[26]. Rather, we offer an account of the psychology of causation and in particular of how we experience and engage with causes in daily life.

Building on the interventionist accounts of causation, psychologists and philosophers have developed an account of the function of causal cognition (Hitchcock, 2017; Hitchcock & Knobe, 2009; Kirfel et al., 2021; Lombrozo, 2010; Vasilyeva et al., 2018; Woodward, 2014). The main thesis of that position is that one central function of judging causes is:

> *to identify relationships that can be exploited for manipulating and controlling the world by intervening on them*

Our proposal is built on this psychological interpretation of the interventionist approach to causation but goes beyond it. Instead of interpreting interventionism as a purely functionalist account, we propose that the act of intervening plays a more intrinsic role in causal cognition rather than functioning as its "goal". Following the enactive view of cognition we take intervening to be an intrinsic aspect of causal cognition in daily life. *What people are doing* when they are engaging causes in their environment is identifying relationships and exploiting them by intervening on them. Hence our approach will be to characterize these phenomena – the process of identifying relationships, the character of these relationships, and controlling the world by interventions - in ecological and enactive

[25] These accounts are also sometimes referred to as 'agency', 'manipulationist', or 'manipulability' theories of causation.

[26] Take for instance the prominent critiques of anthropomorphism and circularity that interventionism has received repeatedly (Woodward, 2016). The charge of anthropomorphism is about the fact that agents are put at the center of defining causation, while causation is a feature of the world independent of agents. The charge of circularity refers to the idea that 'intervention' itself is a causal notion and so cannot be used in an account of causation. Neither of these apply here as we are not offering an account of causation itself but rather of causal engagement.

terms. Doing this will lead us to appreciate the roles that learned abilities, practices, and wider sociocultural context play in determining what we perceive or judge as causal. We will describe the identification of causes as a special instance of selective attention, causal relationships as ecological information, and intervention possibilities as affordances. Let us start with the process of identification.

7.5 IDENTIFICATION OF CAUSES AS SELECTIVE ATTENTION

The psychological process of *identification* as such has received little attention in the literature on causation. What does it mean when we identify something? Our starting point in answering this question (in relation to actual causation) is to look at a necessary condition of identification. When we identify something we necessarily pay *attention* to it. On the E-E account attention should be understood as the *selective openness to relevant parts of the environment* (Chemero, 2003; E. Gibson & Rader, 1979; Rietveld & Kiverstein, 2014). Relevancy here is determined by what matters to the organism, those things that are related to either the improvement or degradation of its situation. This selective openness forms the basis of selective engagement with only those affordances that are relevant. Viewing attention in this way, we can understand the identification of an actual cause as a state in which an agent is selectively engaged with that cause. When we identify something as an actual cause, we engage with that cause and not with other possible causes. By engaging with the identified cause, we are open to the action possibilities (affordances) that it offers in conjunction with our abilities.

It is true that we often judge a single factor to be the actual cause of some event, however we can also judge multiple factors to be causes of that event. So while the identification of events is not strictly exclusive in that we can only pick out one cause, it is at the least *selective,* as we simply cannot engage with all possible aspects of our environment at the same time.

Similar to the amount of possible causes, the amount of affordances in our environment is plentiful (Rietveld & Kiverstein; 2014). This raises the question how we become responsive to only the relevant affordances in a situation. Applied to the topic at hand, this question becomes how we come to identify particular relations or events as causal and not others. This is the problem of causal selection (Hesslow, 1988): why do we pick out only certain causes and not others? Put differently: How are we selective like this?

To answer these questions, we need to see identifying causes as a skill or ability that one develops throughout life (see Noë, 2012). Viewing this as a skill, as something we *do,* allows us to see that we can be better or worse at it (depending on circumstances). To be precise, the skill that we refer to here is the skill to correctly identify something as an actual cause, i.e. to be selectively engaged with only specific events that are concurrently identified as causes. We used the word ‘correctly’ to indicate that there is a type of normativity at play here. The act of judging a cause can be better or worse for an agent. This normative aspect makes that people often agree on what a cause is. For example, if someone told a group of people that “my dog caused a thunderstorm by barking at the sky”, there would (hopefully) be unanimous agreement that she was wrong and it would reflect negatively on her. In this sense the causal judgment is incorrect. This is a type of normativity inherently dependent on context, which has been dubbed *situated normativity* (Rietveld, 2008; Van Den Herik & Rietveld, 2021). We will return to this notion of normativity later.

Construing the activity of making causal judgments as a learned skill makes it clear that investigating the way in which it is learned could help explain the patterns of judgments adults make. For this reason we will look at how we get better at this skill and formulate an ecological account of this development in the next section.

7.5.1 Ontogeny of identifying causes: education of attention

Ecological theories of learning hold that learning is the process by which an individual becomes better adapted to environment they interact with, i.e. they change to fit better in their ecological niche (Araújo & Davids, 2011; E. Gibson & Pick, 2000; J. J. Gibson & Gibson, 1955). We learn to become selectively engaged with only the relevant affordances in our environment through the *education of attention* (E. Gibson & Pick, 2000; J. Gibson, 1966, 1979), which Gibson characterized as “a greater noticing of the critical differences with less noticing of irrelevancies” (1966, pp. 52). Attention here is again understood as the selective openness to affordances that are relevant for the current activities of the agent. For example, when learning to ride a bicycle, we start to better notice the critical differences resulting from pushing or pulling the handlebar, and start to notice less those aspects that are irrelevant for effective cycling (e.g. the shape of the handles on the handlebar).

So the question of causal selection becomes the question of how we become selectively open to certain aspects of the environment, those aspects that we refer to as actual causes. The basis of this process is an individual’s repeated interaction with their environment, which allows them to identify the relevant regularities. For example, crawling through puddles of water can teach an infant that touching water causes their clothes to get wet and cold. In this way learners use the sensorimotor feedback they collect to educate their attention towards the most useful perceptual information (J. J. Gibson & Gibson, 1955; Jacobs & Michaels, 2007).

On top of repeated interactions with the environment, the education of attention can be facilitated by supervision. We highlight this supervision here as it gives us additional clues to the situated and sociocultural nature of causal cognition. Supervising the education of attention is done by skilled individuals who selectively introduce someone to the relevant aspects of the environment and the affordances associated with them

(Ingold, 2001). Skilled individuals (e.g. parents) guide a child's attention towards the specific aspects of the environment. To develop the skill of identifying actual causes, caregivers guide the attention of an infant to a cause when the goal is to manipulate or understand (as a proxy for future interventions) a certain outcome. Such guiding of an infant's attention can be done using linguistic or gestural acts.

Both explicit (linguistic) and implicit (non-linguistic) directions of attention can direct attention to causes. Such directions of attention can be understood as *attentional actions*, that is, recognizable and repeatable forms of behavior performed by one person to indicate an aspect of the current environment to another for some purpose (Van Den Herik, 2018). For example, a parent can point to a puddle of water after seeing that their child is observing their wet clothing and thereby link cause and effect. The important part of this process is that the attention is directed at a specific aspect of the environment (the actual cause of some event). While this is initially directed by a caregiver, ultimately the learner will be able do this later without direction. Repeated experiences of co-occurrences of causes and effects will build up her skill at detecting causes. In this way the learner becomes sensitive to the right parts of the environment, which enables her to execute effective interventions. Hence identifying actual causes is a very basic skill and it being learned partly through non-linguistic attentional actions shows that it is not necessarily linguistic, it can encompass both linguistic and non-linguistic behavior.

7.5.2 Identification of actual causes as skilled causal engagement

That identifying causes is learned through both linguistic and non-linguistic behavior helps us characterize it further. The behaviors we have discussed so far are often described in the literature either as 'making causal judgments' or as 'causal perception', but these might not be the best terms to use. 'Making causal judgments' tends to be associated with explicit reporting of a cause. This is only necessary in experiments, in daily life the situation often requires us just to act after we identify a causal relationship. For example, when a mother sees her baby crying and judges the cause of this to be that she is hungry, no words are necessary for the mother to start breastfeeding. It seems to us that the notion of 'judgment' starts to become strained here, as we seem to be discussing something more general. It is unclear what judgment exactly refers to. Does it refer to the perception, a decision, an act, an utterance, or specific behavior following a specific type of perception? The term judgment seems to come with notions of conscious awareness and the explicit reporting of an experience, both of which need not be the case.

A better term for how we engage with causes would be more descriptive and clearly cover all behaviors described hitherto. What underlies all examples of behavior discussed so far is a type of skilled perception (see Noë, 2012). That is, the ability to attend to and so perceive the relevant aspects of the environment, namely, the actual causes.

However, using the term ‘causal perception’ does not seem intuitive either and would be confusing due to its use in the literature. Certain cases, mostly involving physical causation, tend to be described as causal perception, such as when viewing billiard balls colliding (e.g. Michotte, 1963). Other cases are more naturally described as involving causal judgments and they are also generally thought of as involving “higher” cognition[27]. These cases tend to involve linguistic expressions, such as in experiments using vignettes where participants are asked to rate to what extent certain factors are causes of some event. What we are targeting is something that covers both “lower” and “higher” cognition, as it involves what happens when we look at billiard balls colliding as well as when we reason about causes in a vignette.

Luckily we have no need to provide a distinction between what is perception and what is a judgment, nor between what is traditionally divided as “lower” or “higher” cognition. Since we attempt to characterize something more general, common to both these types of cases, we will use the notion of *skilled causal engagement*. We use the term ‘skilled’ because it is an ability that we need to learn and that we can get better at. We use the term ‘engagement’ as this is the starting point of all the phenomena we discuss. Whether described as ‘perception’ or ‘judgment’, in all these instances an agent is engaged with a particular aspect of the environment, regardless of whether it is followed up by some form of communication, an act/intervention, or further reasoning. Throughout the rest of this paper we will still use the terms ‘judgment’ and ‘perception’ when discussing particular examples where they seem most natural. However, our account does not distinguish between them, and views them both as instances of skilled causal engagement.

We are now able to give an E-E description of the ‘identification’ referred to in the interventionist view of causal cognition. This identification is the selective openness to the relevant aspects of the environment, i.e. those aspects we deem to be actual causes. This openness results in selective engagement: we act only upon those relevant (the ones we have deemed causal) aspects of the environment. Since it is this selective openness manifest in engagement that is crucial in perceptions and judgments of actual causation, we will refer to the phenomenon as skilled causal engagement, which is defined as: the ability to be selectively open to or attentive of relationships that can be exploited for purposes of manipulation and control by intervening on them.

[27] Such causal judgments can be considered as “higher” cognition as they can, for instance, incorporate complex information over an extended time period and can involve environmental aspects not directly present to the senses.

7.6 CAUSALITY AS ECOLOGICAL INFORMATION

7.6.1 Causal regularities

The interventionist view on causal cognition refers to the *identification of relationships*. We have just analyzed the process of identification using the E-E view of cognition. If we view this ‘identification’ as selective openness, what comes of the ‘relationships’? Within the E-E framework the concept of *ecological information* refers to the structures or regularities in the sociomaterial environment encountered by an organism (Bruineberg, Chemero, et al., 2018; J. Gibson, 1979). Causal relationships constitute part of the regularities we encounter in the world. When A causes B, we tend to encounter A and B together in the world. Causal regularities are part of the ecological information through which we are coupled with the environment. Let us take another look at how the interventionist account of causation (Hitchcock, 2017; Hitchcock & Knobe, 2009; Woodward, 2014) characterizes the relationships involved in judgments of causation. It posits that the goal of causal cognition is to:

> *identify relationships that can be exploited for manipulating and controlling the world by intervening on them*

We take this to be true descriptively for much of our causal engagement in daily life. What people are doing when they judge causes is identifying relationships that can be used for interventions. We contend that these two things are the same from a psychological and phenomenological perspective. Those relationships that are exploitable for manipulation and control through interventions are the ones we mostly experience as causal. This statement is not intended to be about the metaphysics, ontology, or epistemology of causality[28]. This is a statement about human psychology. Crucially, we contend that what we typically do when we judge, reason or talk about causes is judging, reasoning, or talking about *relationships that we can or could intervene upon to manipulate the world*.

In most circumstances, when we are looking for the cause of some outcome, we are looking for an aspect of the environment that we can manipulate in order to change the outcome. When we are looking for the cause of our car failing to start, we are looking to fix it. When we are looking for the cause of our glass falling over on a table, we are looking to stop it from falling again. When we are looking for the causes of a successful birthday party we hosted, we might be looking to replicate it again next year. We return to this role of interventions in Section 7.

[28] We are aware that, taken to be true, it might have its consequences for the philosophy of causation, but that is not the topic of this paper.

For now, we can appreciate that relationships that can be exploited for manipulating the world constitute many different regularities that we encounter in the world. In other words, causality is a form of ecological information that allows for manipulation and control. Let us specify this further.

7.6.2 Causal relationships can be both law-like and conventional

Traditionally the focus of research in ecological psychology has been on *lawful* ecological information in order to explain the informational coupling between organism and environment (J. Gibson, 1979; Turvey et al., 1981). The regularities present in lawful ecological information are due to our world being governed by physical laws. For example, there is a lawful relationship between the shapes of objects (as felt by touching them) and the patterns of light they reflect.

Importantly, it has been argued that the information provided by lawful regularities in the environment is not enough to account for the diversity and richness of affordances available to humans (Rietveld & Kiverstein, 2014; Bruineberg, Chemero, & Rietveld, 2018). The key insight here is that for humans, affordances are not just specified by lawful regularities in the environment. On the contrary, the majority of human affordances are at least partly determined by sociocultural practices (Kolvoort & Rietveld, 2022). Most of our actions take place within a context of practices and conventions that have been laid out by others before us.

Bruineberg and colleagues (2018) introduced the notion of *general ecological information*[29] to capture all regularities in the environment that specify the actions possible to humans, conditional on their skills. This notion is defined in an evidential sense as “any regularity in the ecological niche between different aspects of the environment (X and Y) such that the occurrence of X makes Y likely” (Bruineberg, Chemero, et al., 2018). The regularities that fall under lawful ecological information are such that one aspect (e.g. shape) *determines* the other (pattern of reflected light). In contrast, the regularities in general ecological information require only that one aspect of the environment *constrains* another aspect. Like how a label on a cardboard box constrains the likely contents, or how the muffled sounds from a neighbor’s apartment constrain what your neighbors are likely doing. Hence, these type of regularities are also referred to as *conventional* constraints to contrast them with law-like constraints.

[29] Whether general ecological information can fill the role that lawful ecological information does in traditional ecological psychology is still debated. This relates to the question whether ‘conventional constraints’ (instead of ‘law-like constraints’, see below) can allow for the perception of affordances. While these are important debates, they are beyond the scope of this paper and we refer the reader to the literature dealing with this discussion (Bruineberg, Chemero, et al., 2018; Golonka & Wilson, 2019; Turvey et al., 1981; van Dijk & Kiverstein, 2021).

How do causal relationships fit within this conceptual framework? Certainly it is the case that some exploitable relationships can be characterized by one aspect of the environment determining the other, as in law-like ecological information. An illustration: The breaking of a wineglass is determined law-fully by a force acting upon it. Hence, we can say that some force caused the wineglass to break. This is an exploitable relationship, since we can impact the outcome (the wineglass breaking) by intervening on the cause (the force). This provides us with the action possibility of breaking a glass (by putting a force on it) or to stop a glass from breaking (by removing or stopping a force impacting it).

However, it can also be the case that an exploitable relationship is only conventional and not law-like.[30] This happens when one aspect of the environment constrains (but not strictly determines) another aspect of the environment. These relationships are exploitable when the constraint is reliable enough so that it can be adaptive to act upon the constraining aspect to impact the outcome. One example of this is the relationship between emotional states and behavior. We often perceive and make statements about how emotions cause behavior, like "his anger caused him to punch a wall". There is no law-like relationship between anger and aggressive behavior, not every angry person becomes aggressive. There is a conventional regularity here though, emotional states of anger tend to co-occur with aggressive behavior. Even though the relationship is not law-like, our claim is that we perceive the relationship to be causal since in certain situations we are able to stop aggressive behavior from occurring by intervening on someone's emotional state, by calming them down for example. This is what makes us perceive the relationship in those situations as causal.[31]

Causal regularities are a form of general ecological information; both lawful and conventional regularities afford intervening in a way that is adaptive. Conceiving of causal relationships as ecological information highlights that they are inseparable from the affordances available to us. This allows us now to leverage what we know about affordances to understand causal judgments.

[30] That causal relations can also be encountered as conventional regularities is not a novel idea. Existing probabilistic approaches to actual causation already incorporate this idea, in such frameworks causes increase or decrease the probability for the effect to obtain and hence causes do not strictly determine their effects. However, such accounts are not well suited to incorporate abilities and the concrete situation as they are formalized using graphs (i.e. Causal Bayesian Networks) which are limited in representing such contextual factors. In the next sections we will discuss the role of abilities and situational context and argue that they are crucial in understanding causal cognition.

[31] There is a related discussion in the literature on whether reasons for acting can be considered as a cause of the action (see Davidson, 1963; Dretske, 1989). In this article we focus on external causes, i.e. causes that are located in the environment of the agent who perceives a causal relationship. Future efforts could look to expand the ecological-enactive account to also include causes 'internal' to the agent.

7.6.3 Causality: a relational affair involving abilities in context

Humans grow up in highly complex cultures that allow for specialization, we learn very specific skills that distinguish us from others. The education of attention develops differently for all of us and this leads us to be capable of different interventions.

We will illustrate below how being educated to perform specific interventions is related to making different causal judgments, i.e. to differences in skilled causal engagement (Gallagher & Zahavi, 2008; Noë, 2012). But before this it is important to note that we are not arguing for the existence of inter-individual variation in causal judgments. This has been established empirically. Glymour et al. (2010, p. 187), referring to an experiment on actual causation by Walsh and Sloman (2005), aptly recognized that: "Their results were decidedly ambiguous: except in the clearest cases—those on which the entire philosophical community agrees—the modal description for each situation was provided by 60% or fewer of the participants.". It goes beyond the scope of this paper to provide an overview of all of the relevant empirical results on causal cognition, for our purposes it is important to know that the findings of Walsh and Sloman (2005) are not an exception. A lack of unanimous agreement on causal ratings is the norm[32]. The traditional theories have problems with accounting for this variability as dependence and transference considerations shouldn't differ between people. Our account, on the other hand, can explain this variability by appealing to differences in abilities and practices that agents are a part of.

To understand how differences in abilities impact what we experience as causal, we need to take into account that affordances are relative to abilities (Heft, 1989; Noë, 2004; Rietveld & Kiverstein, 2014). With regard to affordances, Kiverstein, van Dijk, and Rietveld (2019) proposed to distinguish between two levels of analysis: the individual and the 'form of life'. Here the term 'form of life' refers to "the relatively stable and regular patterns of activity found among individuals taking part in a practice or a custom" (Kiverstein et al., 2019; Wittgenstein, 1953). The notion of a *field of affordances* refers to the *relevant* action possibilities that are afforded by a specific environment to a specific individual. We can interpret the field of relevant affordances as those aspects of the environment that a particular individual is able and ready to engage with. The notion of *landscape of affordances* is used to refer to available affordances in relation to abilities

[32] For the reader interested in more examples of variation in causal judgements see (Beller et al., 2009; Bender & Beller, 2017; H. Choi & Scholl, 2004; T. F. Icard et al., 2017; Kirfel & Lagnado, 2018; Kominsky et al., 2015; Rehder, 2014; Samland & Waldmann, 2016; Vasilyeva et al., 2018; Walsh & Sloman, 2011). As these studies do not report full response distributions, one can look at the standard deviations of the reported mean judgements as an indication of the substantial inter-individual variation. Note that these works do not study variability itself. One recent study that does specifically target variability in causal judgements reports substantial variability both within and between participants (Kolvoort et al., 2021).

available in a form of life. It is in these different forms of lives, e.g. different sociocultural practices, where different abilities and skills are developed.

Now we can understand how different skills that let us intervene in the world can lead to the experience of different causes (see Gallagher & Zahavi, 2008; Noë, 2012). As an illustration of abilities in the context of different sociocultural practices, let us look at two people, a neurosurgeon and a lawyer, who have a friend that suffers from tremors. The lawyer might judge the cause of these tremors as being a 'medical problem'. The neurosurgeon, however, will likely judge the cause to be different, something more specific, such as a lesion in a particular brain area. This difference arises because in the practices of which the neurosurgeon is part of (i.e. neurosurgery) there are skills available that are not available to lawyers and so they inhabit different landscapes of affordances. Over many years neurosurgeons are trained to attend to very specific aspects of our nervous system in order to intervene in this system. In the form of life of neurosurgeons there are skills available to distinguish between different parts of the brain, these skills are not available in the practices of lawyers. Hence, the fields of relevant affordances are different for the lawyer and the neurosurgeon in the context of this concrete situation, they are solicited by different aspects of the environment (cf. Withagen et al., 2012). An affordance, i.e. a possible intervention, for the lawyer would be to send his friend to the hospital, consistent with his causal perception of a 'medical problem'. The field of relevant affordances in this case is different for the neurosurgeon. In her form of life there is the ability available to operate on the nervous system and she might have specifically encountered ecological information of a form that constrains the type of neurological issues people face when they have tremors. Her being part of this practice has made her skilled causal engagement function in a particular way: she can identify a lesion in a particular brain area as the cause of the tremor. While the lawyer and neurosurgeon would probably agree on what the actual cause is after conversing, their initial identification of the cause of the tremor is different due to their different skills and learned practices.

A similar analysis applies to the car crash example mentioned in the introduction. In the example a policeman, engineer, and a psychologist judge the cause of a car crash to be different (Carnap, 1966). Again, our affordance-based account naturally points us towards the different skills these persons have. Policemen, engineers, and psychologists have been trained in different practices to be sensitive to different parts of the environment. This has formed their skilled causal engagement. The policeman judged the cause to be the driver's speeding as he has learned to intervene on this by writing speeding tickets. The engineer judged the road to be the cause, an object he could modify or repair. And similarly the psychologist focused on the driver's mental state, as mental states are where she has learned to intervene.

Our affordance-based approach helps understand the situated causal selection problem by appealing to the available skills and relevant social, cultural, and material practices. In this way it can understand why different people perceive different causes, something existing accounts struggle with. We simply cannot reduce the problem by appealing to a single criterion (Lombrozo, 2010) such as statistical dependence, transference of force, or even the quality of an explanation that the cause might provide. However, this does not mean causal judgments are completely subjective or that they cannot be incorrect. The phenomenon of situated normativity discussed in the next sub-section will help to see this.

7.6.4 Situated normativity and objectivity

There is a clear normative dimension to the things we do embedded in the practices we are part of. This is captured by the notion of *situated normativity* (Klaassen et al., 2010; Rietveld, 2008; Van Den Herik & Rietveld, 2021), which refers to the normative aspect of cognition in skillful action. This notion implies "distinguishing adequate from inadequate, correct from incorrect, or better from worse in the context of a particular situation." (Rietveld, 2008). Situated normativity is what makes an individual's actions adequate or not. In every concrete situation an individual distinguishes between better or worse actions. Whether some action is adequate or not is dependent in part upon agreement among members of a sociocultural practice.

Let us continue the previous example concerning the neurosurgeon and the lawyer to illustrate this. Abstracting away from context, neither the judgment that the cause of the tremor is a 'medical condition' nor that the cause is 'a lesion in a particular brain area' is wrong. In a way both are right and neither proves the other incorrect. This is different when we look from *within the context of a practice,* which is where we find a strong sense of normativity.

Within the practice of neurosurgery, the practitioners have a clear sense of what is right and what is wrong. Claiming the cause of a patient's tremor to be 'a medical condition' does not agree with the standards and patterns of behavior that are the norm within the field of neurology. One can easily imagine that such a claim is frowned upon in a meeting of neurosurgeons.

This example illustrates that judgments of causation form a part of human practices. Practices differ in what causal judgments they allow for, which is dependent on the type of interventions they tend to engage in. Within these practices the situated normativity imbues actual causation with a type of objectivity, what we will refer to as the *situated objectivity* of skilled causal engagement.

7.6.5 Causal engagement spans over the objective-subjective and material-social dichotomies

We just discussed differences in abilities or skills as a source of variation in causal judgments. The complement source of variation lies in the environment. While the physical laws responsible for law-full regularities are the same for everyone, the sociocultural practices giving rise to conventional regularities differ from one culture to the next. As discussed earlier these conventional regularities impact what we experience as causal. Since these conventional regularities and their relevancy depend on cultures and practices, people, by virtue of being part of different cultures and engaging in different practices, will perceive causality as pertaining to different regularities[33].

Taken together, differences in skilled causal engagement, due to the fact that the education of attention is idiosyncratic, can explain differences in causal judgments (i.e. identifying causes) between individuals in a culture or within a sociocultural practice. In addition, differences in the conventional regularities encountered in the world can explain the variance of causal judgments between cultures and individuals part of different sociocultural practices. While we can distinguish these two sources of variation on theoretical grounds, in reality they are of course strongly intertwined as the skills available in a form of life depend on the environment and vice versa. Ultimately, this variation in people's judgments of actual causation underlines that the psychological reality of causality as ecological information is situated and relational: it connects people's skills with their environment, the causal information we engage with constitutes a relationship between us and the environments we inhabit.

7.7 INTERVENTIONS AS ENGAGING WITH RELEVANT AFFORDANCES

We have now analyzed the process of identification and the relationships involved in causal judgments from an E-E perspective. What still needs to be unpacked are the interventions that can be executed when engaging with causal regularities.

According to the interventionist theory of causality, causes can be viewed as "handles for manipulating or controlling their effects" (Woodward, 2011, pp. 8)[34]. While literal handles mostly just afford grabbing, the figurative handles Woodward refers to afford a lot more. Causal relationships, the identification of them and the acting upon them, are ubiquitous in (human) life and so there are many types of actions that causal relationships afford us. To characterize such actions and their surrounding dynamics we need to look at the whole organism-environment system and at what drives an organism to act. For this it is helpful to use a running example:

> EX2. A man sitting in a cafe sees his glass slowly move over the table and grabs it to stop it from moving further. Looking at the surface of the table he notices it is not completely horizontal. He puts one hand on the side of the table and pushes down, the table pivots somewhat and is now slanted towards the other direction. He pushes on the other side and sees the table wobble to

[33] Cross-cultural studies on causal judgement are rare, noteworthy exceptions are (Bender & Beller, 2011; I. Choi et al., 1999; McGill, 1995). These studies all provide evidence for significant cross-cultural variation in causal judgements.

[34] A very apt metaphor for an affordance-based account, as there is empirical evidence for literal handles evoking affordance effects (Tipper et al., 2006).

> its original position. Looking underneath the table the man sees that one of the four legs of the table is not touching the floor. He promptly grabs a few coasters from the table, puts them underneath the suspended table leg. This stabilizes and levels the table making sure that the glass will not fall off.

Let us first regard the skilled causal engagement and ecological information contained in this example, after which we will turn to the interventions involved and see how we can characterize them.

7.7.1 Skilled causal engagment and ecological information as basis for interventions

The man first perceives that the glass is moving, then he selectively attends to the table, which prompts him to attend to the table legs, and this ultimately leads him to put coasters under one of the legs. His attention flows from one relevant aspect of the environment to the next, from glass to tabletop, from tabletop to the table's legs, and from there to the coasters. This is skilled causal engagement. The man in this example identifies the following causal chain:

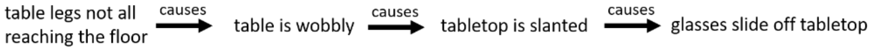


table legs not all reaching the floor
causes
table is wobbly
causes
tabletop is slanted
causes
glasses slide off tabletop

Figure 7.1 *Perceived causal relationships in EX2.*

Note that the man observes the elements in this chain in reverse, he starts by observing the glass sliding off the table. Subsequently his attention is repeatedly guided from an effect to its cause. The behavior of the man would be impossible without a sense of the causal relationships involved. That the man perceived this causal chain is due to his skill in causal engagement. It is an example of skilled behavior, the whole sequence can play out in under half a minute and someone without experience with tables and glasses would have a hard time replicating that feat. As discussed earlier, skilled causal engagement is the selective openness to relevant relationships in the environment that allow for effective interventions. It is this selective openness that leads the man from one relevant aspect of the environment to another, and so leads him to quickly stop his glass from repeatedly falling off the table.

The ecological information that formed the basis for the education of attention that enabled identifying the causal chain above consists out of co-occurrences of sliding glasses and slanted tabletops, of slanted table tops and wobbly tables, of wobbly tables and not all table legs touching the ground. Via previous co-occurrences of any of the above events with the event of someone using coasters to level a table, the man was educated to attend to nearby coasters (which in turn was made possible by the conventional regularity of cafes having coasters). His use of a coaster to level the table is an intervention in the causal chain that led to his glass sliding off the table. This can be represented as follows:

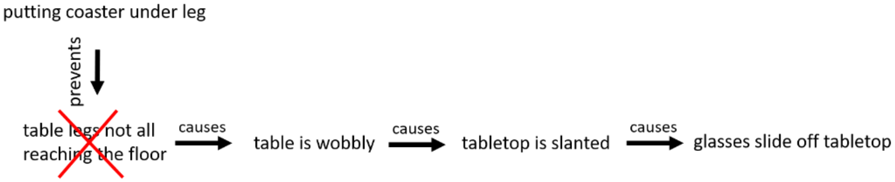

putting coaster under leg
prevents
table legs not all reaching the floor
causes
table is wobbly
causes
tabletop is slanted
causes
glasses slide off tabletop

Figure 7.2 *Perceived causal relationships after intervention in EX2.*

7.7.2 Relevant causal affordances are intervention possibilities that improve grip

By intervening in the causal chain the man in our example stops the ‘effect’, which is the glass sliding from the table, from occurring. This intervention would not have been possible without his identification of this causal chain. In this manner skilled causal engagement helps to increase a person’s *grip* on a particular situation, in other words, it improves one’s *grip on the field of affordances*. The notion of *tendency towards optimal grip* (on the field of relevant affordances) describes the basic concern of any organism to improve its situation (Bruineberg & Rietveld, 2014).

This tendency is closely related to the notion of situated normativity: where situated normativity denotes that there are better and worse actions in a certain context, the tendency towards optimal grip involves executing those actions that are better, that is, to deal adequately with the field of multiple relevant affordances. The interventionist credo involved relationships that can be *exploited for purposes of manipulation and control*. On the E-E account the manipulation and control referred to here are ways of improving grip on the situation.

Now we can ask ourselves: what led the man down this path of perceptions and actions? It is those affordances that will improve an individual’s grip on a particular situation that solicit that individual’s actions (Dreyfus & Kelly, 2007; Rietveld, 2012; Withagen et al., 2012) and those soliciting affordances are perceived. These soliciting affordances are the *relevant intervention possibilities*.

In our example, the man experiences directed discontent towards the glass falling off the table. Directed discontent is a phenomenological expression of situated normativity (Rietveld, 2008), it is what is experienced or felt in a situation that can be improved. The man experiences the glass staying on the table as being better than it sliding off the table. This is the point at which different people’s behavior would diverge based upon their skilled causal engagement, i.e. their abilities. While the intervention possibility of stabilizing the table using a coaster is always present, only those with the necessary skilled causal engagement would have their attention guided in such a way to be able to act upon this affordance (Noë, 2012). People who do not have this skill might engage with a different affordance, like that of asking a waiter for a different table. Both these actions are effective interventions in the causal system described by the example, effective in the sense that they lead to improved grip on the situation, which here means having a glass that does not slide off the table.

Ultimately, what led the man down the particular causal chain involving the table and its legs was the possibility of an effective intervention, that is, a relevant affordance. Without the possibility of this effective intervention the man would not have selectively engaged with this causal chain, nor would there be any reason to. We contend that causes are generally identified as such in virtue of the possibility of an effective intervention (see next sub-section).

7.7.3 Is it nothing but intervention possibilities?

Our thinking is in line with the idea that causal judgments and explanations are used for the identification of relevant interventions, which has been proposed before (Hitchcock, 2012, 2017; Hitchcock & Knobe, 2009; Kirfel et al., 2021; Lombrozo, 2010; Vasilyeva et al., 2018). However, our account goes further. We directly relate the experience of causality to possible interventions (relevant affordances), meaning that we contend that we are directly sensitive to relevant intervention possibilities as these solicit action (Dreyfus & Kelly, 2007; Rietveld, 2012; Withagen et al., 2012). Existing accounts posit that we are sensitive to particular dependence and transference considerations – such as stability, normality, and portability – and that these considerations in turn guide us towards effective interventions (Hitchcock & Knobe, 2009; Lombrozo, 2010; Vasilyeva et al., 2018). Additionally, we contend that we are sensitive to intervention possibilities (relevant affordances), which in turn guide us towards environmental regularities that are stable, normal, and portable. Information that provides clues to intervention effectiveness (such as the stability of a dependence relationship) is relevant for how we experience causality mainly because they are clues to intervention possibilities, not because they have some inherent value. For instance, there is a very stable relationship between the presence of oxygen and forest fires, however oxygen itself does not provide an opportunity to intervene and so people do not tend to rate it as causal (Cheng & Novick, 1991). Instead, people tend to rate a less statistically normal factor, such as the lighting of a match, as causal. In this case the consideration of normality trumps that of stability (not every lit match results in a forest fire). Which considerations are important is determined by the possibilities of intervention in the particular context. Putting these intervention possibilities as affordances at the forefront of our account of how we engage with causality, makes it so that we can assign the proper relevance to factors that have been hitherto underappreciated: learned skills and the wider sociocultural practice in which causal cognition takes place.

Our earlier statement that people *generally* identify causes in virtue of intervention possibilities needs some qualification. This is not to say that a direct intervention is always possible. Rather the idea is that in everyday life the identification of causes, either in form of a perception or judgment, primarily involves identifying intervention opportunities. There are of course exceptions. We can learn about causal relationships not by being in direct contact with them, but through communication with others. And it might be the case that a particular relevant intervention was available to someone at a different time or place, but not anymore at the time and place where this information is communicated. Future research could aim at developing an affordance-based account of such dynamics across time and space. Other examples of causal claims that tend not to be related to intervention possibilities are those that involve deities or supernatural phenomena. We believe that these type of instances are exceptions to the rule. People can, for a variety of reasons, express that anything is causal. However, we believe that in most instances that we engage with causation in our daily lives, there is a relevant affordance present as well.

In these concrete situations relevant affordances play a principal role. However, concrete situations are often more complex than meets the eye and so they require scrutiny of the particulars to understand the affordance dynamics at play. We can illustrate this by looking at the complexities at play in EX2, which also illustrates the constraints of the prominent psychological theories.

7.7.4 Concrete situations are complex and so is causal selection

The standard psychological theories of causal reasoning, using either difference-making or transference criterions, do not provide much guidance in a concrete situation like EX2. In EX2 There are too many factors that are connected in one way or the other to the glass falling of the table. For example, the smoothness of the table and the shape of the glass are on these accounts also causes. Then why are they not selected? These factors do not allow for effective intervention and so they play no role for the agent in our example. While people can study the table surface and the shape of the glass such that they will be identified as causes, most likely they will not. And this is exactly what we would like to explain. Here we see that the notions of improving grip and possibilities for interventions allow the affordance-based account to be more selective and alleviate the problem of underdetermination of existing accounts of causal selection.

It is important to realize that the processes we have described are part of the vastly complex dynamics between agent and environment. One source of such complexity is the fact that the field of relevant affordances is ever changing.

We can find an illustration of this complexity in our running example. In the example, after noticing the table was slanted, the man pushed down on one side of the table and discovered it was unstable. Pushing the table became a relevant affordance after the man perceived that the tabletop was not levelled. In itself this action can be construed as an instance of skilled causal engagement: after finding out that the table was slanted, the man, through an intervention, identified that the cause of this was an instability of the table (cf. Gallagher, 2017; Noë, 2004). Pushing on the table is an action, but it is also crucial in identifying that the table was unstable. In this way an intervention can enable the discovery of another affordance, i.e. interacting with causes can further the education of attention. Moreover, that the table was slanted afforded pushing on its corners to test its stability, the instability afforded improving grip by placing coasters under the table legs. Hence the affordance of pushing on the table was nested within the affordance of stabilizing the table.

This is not an exceptional case, to the contrary, we are generally engaging with a multitude of relevant affordances over different timescales simultaneously (Kolvoort & Rietveld, 2022; Rietveld, 2012; van Dijk & Rietveld, 2021a). Situations unfold continuously and we deal with this in a similarly continuous fashion using a multitude of causal handles to help us along the way.

7.8 SUMMARY AND CONCLUDING REMARKS

The interventionist theory of causality views causes as 'handles' that can be used to manipulate the world (Woodward, 2011). In the way literal handles afford grabbing, causes as handles afford intervening. We can think of the tendency to move towards optimal grip on the field of relevant affordances as including a tendency to grab the right causal handles. Hence, our E-E account of the interventionist view on causal cognition involves:

> *Selectively attending to the relevant ecological information in order to engage with action possibilities, determined jointly by individual abilities and the sociomaterial environment, to improve grip on the field of affordances by way of interventions*

This account emphasizes the ecological and situated nature of causal judgments. We have argued to see the identification of causes as an instance of selective attention to particular aspects of the environment which we can understand as a skill: *skilled causal engagement*. This is a lifelong skill developed through the education of attention that results from repeated interactions with environmental regularities, which can be (partially) supervised by caregivers. This skilled causal engagement encompasses both so-called "lower" and "higher" cognition as it describes, for instance, the viewing of colliding billiard balls as well as reasoned judgements about past events.

Next, we construed an account of those causal regularities in the terms of general ecological information. Causality is form of ecological information which we encounter in both law-like and conventional regularities. This has implications for the psychological reality of causality, which we should see as a relational affair between aspects of both the agent and the environment. An agent's skills and the practices they inhabit determine what is causal to them and the particular situated objectivity at play.

Ultimately this makes us understand the notion of effective interventions in terms of engaging with relevant affordances. Effective intervention possibilities are relevant affordances for a person in their particular situation. And the basis of such concrete intervention possibilities are skilled causal engagement and causal ecological information.

This E-E affordance-based account of causal perceptions and judgments provides a unified theoretical framework for understanding how and why we experience causation. By restricting themselves to one objective core criterion (such as dependence or transference), traditional theories of causal cognition apply only to a thin slice of behavior (Danks, 2017; Glymour et al., 2010; Lombrozo, 2010) and fail to grasp the situated and enacted nature of causality in daily life.

However, the affordance-based account provided here is not necessarily at odds with the difference-making and physical transference accounts that dominate current psychological perspectives, but rather it describes causation and the psychological role it plays at a more fundamental level. Our account shows that difference making and transference by themselves cannot fully explain our experience of causality and how we make causal judgments. Many more things factor into what a relevant affordance is – aspects of the environment, sociocultural practices, skills of the individual -, and dependence and transference considerations do not take these into account. We need to accept this complexity of (actual) causation for the human mind and not falsely reduce it to a low dimensional problem.

Our account does justice to the fact that cognition is inseparable from perception, action, and the environment in which it takes place. This view foregrounds the role of concrete actions, skills, and context in determining what we experience as causal. To properly understand the role of causality in the mind we recommend that future research into causal cognition explicitly incorporates sociocultural context, skills, and concrete possibilities for action.

8 General Discussion

With this thesis my aim has been to further our understanding of the causal mind. During this process I tried to practice discipline agnosticism; focusing on interesting phenomena, using methods pragmatically, without taking into account disciplinary boundaries. While I consider this pluralist and pragmatist approach laudable it would be grandiose to say that I succeeded in being such an objective scientist. This thesis is thoroughly colored by my academic interests and my formal education, both of which have been partially shaped by disciplinary boundaries. This can be seen from the main topics in each of the three main parts in this thesis.

In Part 1 designed and ran experiments on causal cognition using experimental techniques from adjacent fields in psychology and cognitive science. Before starting this thesis, I was already familiar with experimental methods using time pressure, response times, and repeated-measures designs. To my surprise, when familiarizing myself with the literature on causal cognition at the start of my PhD I found that these techniques were not yet used to their fullest extent, despite their potential to address outstanding questions. In Part 2 I present two modeling studies in which I develop and test the Bayesian Mutation Sampler. This use of computational cognitive modeling reflects my belief in the potential and necessity of using mathematics to understand the human mind. In the development of part 3 I gave the skeptic in myself free reign and considered ways in which the larger frameworks of psychology and cognitive science fall short in comprehending important facets of the mind. Although it was challenging to diverge from my academic training in this manner, my natural tendency to turn conversations into philosophical debates assisted me in this endeavor.

I will now give an overview of each part, focusing on the results and ways in which my work suggests and paves ways for future research. In the remainder of the discussion, I propose a speculative synthesis of two theoretical frameworks used in this thesis, namely the sampling approach to cognition and embodied cognition.

8.1 Experimental studies

Part 1 presents two sets of experiments on probabilistic causal reasoning. In both, participants are taught information about a causal network system, after which they are asked to infer the state of certain causal variables conditional on other variables in the network. The rationale behind this setup is that it makes participants draw on their understanding of the causal network to make an accurate judgment.

In Chapter 2 I tested the effects of time pressure on causal inference with the purpose of elucidating the cognitive mechanisms underlying causal reasoning. To implement time pressure, we asked participants to draw causal inferences and manipulated the available time to respond while measuring response times. This led to multiple novel findings. We found that time pressure leads to quicker and less accurate causal inferences, in line with findings on other types of reasoning and decision-making (Bogacz, Wagenmakers, et al., 2010; Heitz, 2014). In this study we paid particular attention to systematic patterns of non-normative responding identified in the literature (i.e. deviations from CBN predictions), as these 'reasoning errors' can shed light on the cognitive mechanisms used to respond (e.g. Kruis et al., 2020). We found that participants displayed increased conservatism under time pressure, and that this conservatism was related to participant's lack of confidence in their answers. However, we did not find Markov violations to be affected by time pressure. This was surprising as most theories of causal reasoning would predict Markov violations to increase under time pressure. Specifically, standard readings of sampling-based theories (e.g. the Bayesian Mutation Sampler) and heuristic explanations of Markov violations imply that these violations would increase in magnitude. Together, the finding that time pressure impacts certain response patterns but not others, indicates that causal inferences (and errors therein) are not the result of a single cognitive mechanism. Instead, the underlying processes are likely to be more complex and I suggest that it is probably futile to attempt to model all response patterns using a single cognitive mechanism. It is likely that using mixture modeling or a model incorporating multiple mechanisms to account for the different patterns is more fruitful. The data in this chapter suggest a way to capture the pattern of conservative inferences. As conservatism is affected by time pressure (in contrast to Markov violations) and it is related to participant confidence, it might be that it is the result of a more general phenomenon related to uncertainty, and not a phenomenon specific to causal reasoning (such as Markov violations). I hypothesized that the observed conservatism is due to participants' use of prior information (this is tested in Part 2). This would explain the effect of time pressure on conservatism as in the case of high time pressure participants would have to lean on their prior more than on the evidence they can accumulate during stimulus presentation. This could explain the increase in conservatism as an uninformative prior would push judgements towards 50% (see Chapter 4). Such use of prior information could explain the relation to confidence as well since participants would be aware that they are leaning on their prior and not on evidence gleaned from the stimulus.

The findings from Chapter 2 indicate that causal reasoning is likely the result of one or more complex cognitive mechanisms. This is further evidenced by our results in Chapter 3. In Chapter 3 I present an experiment designed to explore the variability in causal inferences as variability in responses has been used to help constrain theoretical development in adjacent fields. While multiple studies in the literature had commented on the existence of such variability in causal judgments (Rehder, 2018; Rottman & Hastie, 2016), it had not been explicitly studied and the literature at large continued with studying averaged responses. Hence, I designed an experiment to elicit repeated causal inferences from participants. The results, for the first time, showed that the variability previously observed is due to both between- and within-participant variability in responses. Moreover, we established that the within-participant variability is affected by the type of inference presented to participants. For inferences where the state of all unqueried variables in the network were known (e.g. for the network $X_1 \leftarrow Y \rightarrow X_2$, see Figure 1.1, the inference $P(X_1 = 1|Y = 1, X_2 = 0)$), participants were more variable in their responses compared to inferences where not all unqueried variables were known (e.g. $P(X_1 = 1|Y = 1)$). The important implication from this finding is that variability in causal judgments, at least partly, reflects variation in the judgment process rather than just noise. This means that the tradition to focus only on averaged behavior has led researchers away from valuable information. These results form a strong argument that theories just describing averaged behavior are limited in the extent they can help us understand causal cognition. Instead, we should also take into account (aspects of) distributions of responses.

Naturally, the experiments presented in Part 1 have their limitations. Some of these stem from the fact that these experiments employed novel methods to target novel behavioral phenomena and as such were partly exploratory in nature. One such limitation is that I did not establish the extent to which the findings generalize. My experiments only tested a limited number of participants in a limited range of (experimental) contexts all while the patterns of behavior are rather complex. To establish the generalizability of my findings, and generally to further our understanding of causal cognition, we need more experimental studies to test the effects of time pressure and to investigate variability. Such studies would, preferably, conduct similar analyses while varying the domains of study. For instance, we focused exclusively on probabilistic causal inference, but similar experimental methods can be used to study other aspects of causal cognition, such as syllogistic reasoning, causal structure learning, causal-based categorization, or interventions. The combined findings of such studies with the ones presented here could shed light on whether the cognitive processes of interest extend more generally to other reasoning and decision-making domains. Or, conversely, they might establish that the relevant cognitive processes are specific to causal reasoning, or even vary within different domains of causal reasoning. Especially if the suggested experimental work is done in conjunction with computational cognitive modeling efforts, I expect it to accelerate our understanding of the causal mind.

8.2 COMPUTATIONAL COGNITIVE MODELING

In Part 2 I apply computational cognitive modelling to the data from the experiments in Part 1. In this part I mainly focus on developing and testing the Bayesian Mutation Sampler (BMS) and on accounting for the variability in causal judgments observed in Chapters 2 and 3.

In Chapter 4 I established that the Mutation Sampler (MS; Davis & Rehder, 2020), while providing a good account of averaged judgments, fails at accounting for distributions of responses. As a process-level model the MS should be able to predict distributions. In particular, I identified that under a range of reasonable parameter values the model predicts a substantial number of extreme responses (i.e. responses near 0% and 100%), but people tend not to make such extreme judgments. Additionally, the MS did not have a mechanism by which it could integrate prior information, something people clearly do when they make judgments.

These observations indicated that the MS required further development. I did this by incorporating a mechanism into it that integrates the information gained from sampling with a generic prior, leading to the Bayesian Mutation Sampler (BMS). I fitted the MS and BMS to the data from Chapter 2 and found that the BMS outperforms the MS considerably. The BMS predicts mean responses better, but it is in the prediction of response distributions that the improvements over the MS were particularly striking. The BMS predicts the location of the main modes of responses more accurately and it does not predict participants to make extreme responses when they do not.

In Chapter 5 I tested multiple possible explanations of the variability observed in causal judgments. This involved fitting multiple models, including the BMS, to the repeated-measures data from Chapter 3. In terms of quantitative fit the BMS outperformed all other models. However, besides quantitative fit, we also assessed whether the candidate models could explain particular qualitative patterns in the data (Palminteri et al., 2017). Again, here the BMS outperformed the other models, predicting relevant patterns in conservatism, Markov violations, and clusters of responses at 50%. Taking the findings from Chapters 4 and 5 together, the BMS currently provides the best account of causal reasoning that exists in the literature and the stochastic sampling mechanism posited by the BMS is thus a good candidate source of the variability we observe in causal judgments. This is in line with findings indicating that a stochastic sampling mechanism is responsible for the variability in causal learning tasks (Bonawitz, Denison, Griffiths, et al., 2014; Bramley et al., 2017; Denison et al., 2013)

While the BMS outperformed other candidate models in Chapters 4 and 5, it was not able to capture all aspects of behavior. Particularly, for both datasets it failed to account for how variability varies by inference type. This is not a weakness of the BMS alone, all other models fared equally or worse on this aspect of the data. This illustrates that the study of variability in causal reasoning is still in its infancy; the studies presented in this thesis are the first to systematically study this phenomenon. Hence, I recommend future research to continue this line of investigation and use cognitive modeling to improve our understanding of causal reasoning and the variability therein. My analyses suggest multiple concrete ways of doing this. One way would be to let the chain length parameter of the BMS vary over inference types. It seems plausible that reasoners generate more or fewer samples (in other words, engage more or less in effortful thinking) based on the problem that they are faced with. It might be that reasoners use the perceived complexity of an inference type to adjust the amount they sample (Zhu et al., 2020). One way they could do this is by making a judgment on the complexity of a problem before they start the sampling process. Conversely, it could be that they judge the benefits of further reasoning during the sampling process itself. Researchers in a different domain have already proposed an adaptive scheme by which the costs and benefits of generating more samples are weighed on the fly to determine how many samples to generate (Gershman & Goodman, 2014; Hertwig & Pleskac, 2010; Vul et al., 2014). Lastly, my results point towards the idea that causal judgments are generated using multiple underlying processes. It might therefore be necessary to conduct more extensive mixture modelling, possibly incorporating more general psychological processes or heuristics, to improve our understanding of causal reasoning. Any of these projects seem like fruitful ways to extend the BMS and the study of cognitive mechanisms underlying causal reasoning.

8.3 EMBODIED COGNITIVE SCIENCE

In Part 3 I took a radical turn and moved away from traditional cognitive psychology and its use of an information processing metaphor to understand the mind. In Chapter 6 I provided an overview of the Skilled Intentionality Framework (SIF; van Dijk & Rietveld, 2017) and the role that affordances, i.e. *possibilities for action*, play in it. SIF is a philosophical approach that combines insights from ecological psychology and enactivism to understand the embodied and situated mind. It follows ecological psychology and enactivism in conceiving of affordances as central to understanding cognition. SIF construes affordances as relations between an animal's abilities and its environment. Using such a broad definition of affordances allows us to understand any type of skilled behavior in terms of engagement with affordances (Rietveld & Kiverstein, 2014). In particular, we can understand the role of causality in our minds in terms of engagement with affordances, which I flesh out in the next chapter.

In Chapter 7 I provide an affordance-based account of causal engagement. I used the term 'engagement' and not causal 'judgment' or 'perception' to highlight the generality of the phenomenon I am considering and to emphasize the *interaction with the environment*. Causal engagement is a particular type of engagement with the world that underlies both causal reasoning and perception as it occurs naturally. In daily life, causal engagement is about engaging with pathways for successful interventions in the world. This is not to say that a direct intervention is always possible. Rather, the idea is that in everyday life the identification of causes, either in form of a perception or judgment, primarily involves identifying intervention opportunities. There are of course exceptions. We can learn about causal relationships not by being in direct contact with them, but through communication with others. In that case, it might be that a particular relevant intervention was available to someone at a different time or place, but not anymore at the time and place where this information is communicated.

At the core of my account is that *causal engagement is a skill* and this skill is about selectively attending to aspects in our environment that allow for effective interventions. These possibilities for effective interventions are understood as relevant affordances. Which actions are effective interventions (or: which affordances are relevant) depends on the material and sociocultural environment. For example, taking off your shirt might be effective if you are overheating while playing tennis, but is probably less so if you are walking under the desert sun or attending a board meeting.

Construing causal engagement this way allows us to understand the variation in causal judgments between different cultures and between people part of different practices, as being due to differences in skills, practices, and culture. Interventions that are used in one practice might not be in another, and so people part of different practices are likely to experience causality in different aspects of the environment. This is illustrated by a famous example from Carnap (1966) about a car crash. In the example a policeman, a road engineer, and a psychologist visit the scene of the car crash and they all make different judgments concerning the cause of the crash. The policeman is likely to say the cause was the driver's speeding, while an engineer would probably point out the state of the road, and the psychologist the mental state of the driver. I argue this is due to these individuals being part of different practices in which they have developed their skill in causal engagement to intervene onto different aspects of the environment. The policeman intervenes on people's speeding (by writing tickets), the engineer intervenes on the road (e.g. by filling potholes), and the psychologist intervenes on mental states (by therapy).

This account of causal engagement provides a theoretical framework for understanding how and why we experience causation. It has a broader scope than the traditional conceptual framework of cognitive psychology used in Chapters 2-5, which conceives of causal cognition primarily in terms of the processing of statistical information. The sole use of such an information processing metaphor prohibits grasping the embodied, situated and, enacted nature of how we deal with causality in daily life. However, I believe my affordance-based account is not at odds with the traditional cognitive psychology view, but encompasses it, and describes causality and its role in cognition at a more fundamental level. The narrower view used in Chapters 2-5 is warranted as our capacity to use statistical information and judge probabilities in the context of causal structures is immensely impactful and worthy of study on its own. However, we need to keep in mind that this is but one perspective on cognition and that there are more ways to understand the role causality plays for the mind.

My aim with developing an affordance-based account of causal engagement was to construct a theoretical framework that helps us understand causal cognition more holistically. It was not of primary interest to relate this theoretical framework to empirical observations, but it is supported by several empirical findings. Multiple experimental studies have shown that sociocultural context is incorporated into causal judgments (Bender, 2020; Bender et al., 2017; Bender & Beller, 2019; I. Choi et al., 1999; McGill, 1995; Morris et al., 1995) and developmental evidence indicates that the development of causal cognition and concrete motor abilities are strongly intertwined (Muentener & Bonawitz, 2017a; Sommerville et al., 2005). These findings provide indirect support for my embodied account of causal engagement. More empirical research should be done to validate my account and we can derive several predictions from it that allow for such investigations.

My account foregrounds the role of sociocultural context, skills, and concrete possibilities for action in what we experience as causal. Hence, I recommend future research into causal cognition to explicitly incorporate these aspects in psychological studies. Sociocultural context and skills are harder to experimentally manipulate, but their effects can be studied by comparing people with different backgrounds (which can be done easily these days using web-based experiments). Moreover, concrete possibilities for action can be manipulated and my affordance-based account makes a precise prediction that can be tested: all things being equal, people perceive aspects of the environment to be more causal when they can perform (or have experience with performing) an action that changes that aspect of the environment. Over the duration of an experiment participants can be taught that they have control over certain aspects of an artificial environment but not over others. My account predicts that those aspects that participants have experience with manipulating will be judged as more causal. To make this concrete, lets imagine an experiment where participants have to rate to what extent a factor is the cause of some event D in the following causal chain: A → B → C → D. Based on previous research we can expect that participants rate factors closer to D as being more of a cause of D (such that C > B > A; Hilton et al., 2010). In addition, my affordance-based account would predict that if participants have experience with being able to intervene on A, B, or C then the causal ratings for that variable will be boosted. Providing participants with experience regarding such interventions is a methodological challenge as one needs to control for other psychological affects related to attention and familiarity. But it seems to me possible in principle, for example by letting participants interact with certain factors but not let them manipulate it effectively. I see multiple ways of designing an experiment that would elicit the ‘affordance effect’ on causal judgments. One option could be to use a physics simulator (as in e.g. Bramley, Gerstenberg, Tenenbaum, et al., 2018; Gerstenberg et al., 2021) in which participants are able to manipulate certain objects but not others. Another option would be to design an interactive story which participants can make decisions regarding certain aspects of the story but not others. Such research would be able to test the affordance-based account of causal engagement and possibly refine it.

8.4 Speculations on a synthesis of cognitive psychology and affordances

In this thesis I have broadly argued for two ways of understanding causal cognition. In Chapters 4 and 5 I have argued for a sampling approach to understand causal reasoning. I implemented this approach in the Bayesian Mutation Sampler. In Chapter 7 I have argued for an affordance-based view of causal engagement. This naturally raises the question: can we reconcile the theoretical commitments underlying these two approaches? I will not be able to provide a complete answer to this question here (to do that will likely require me to write another thesis), but I would like to point out some common ground. What these frameworks share is that they propose that our minds make use of the concrete world, in some sense they are both *situated*. According to the sampling approach (and BMS) we sample *states of the environment* from a generative model. According to embodied cognition theory (and my affordance-based view), an agent *is* a generative model of their environment (Bruineberg, Kiverstein, et al., 2018) and affordances form a relation between agent and environment. Put very simply, both accounts emphasize that cognition *depends on the environment*, albeit in very different ways relating to their different use of the concept of a 'generative model'. This seems to me to be a possible starting point for research into integrating these theoretical positions. To this end, let me first briefly sketch the way in which the notion of a generative model is used in these respective frameworks, after which I will discuss a possibility for cross-fertilization.

At its core, the sampling approach to cognition posits that we approximate Bayesian inference by way of sampling (Bramley et al., 2017; Chater et al., 2020; Dasgupta et al., 2017; Davis & Rehder, 2020; Denison et al., 2013; T. Icard, 2016; Sanborn & Chater, 2016; Vul et al., 2014; Zhu et al., 2020). As Bayesian computations often require vast computational resources, statisticians and computer scientists have developed methods, such as Markov Chain Monte-Carlo sampling, to approximate these calculations based on taking samples from a posterior distribution (e.g. Hastings, 1970; van Ravenzwaaij et al., 2018). The BMS posits that, when reasoning about a causal system, people think about concrete cases and it is these cases that are the samples for the inference process (Davis & Rehder, 2020). These samples are obtained either by retrieval from memory or by generating them from an internal generative model. Let us focus on the latter.

An internal generative model allows for simulating certain relevant aspects of the world in a manner that allows for useful inferences (Lake et al., 2017). Hence, these type of models are also referred to as simulation models. These theories assume that knowledge about the world is *represented* as a generative model that captures the causal relationships that produce *relevant* outcomes (Bramley et al., 2017; Gerstenberg et al., 2021; Goodman et al., 2015; Lake et al., 2017). These theories (implicitly) follow the structural representation paradigm, which holds that by virtue of mimicking the structure of the surrounding world the generative model represents it (Kiefer & Hohwy, 2019).

The construct of a (causal probabilistic) generative model in theories in psychology and cognitive science is rooted in predictive processing theories of brain functioning and the free energy principle (Clark, 2013; Friston, 2010; Hohwy, 2013). In essence, these predictive processing theories propose that the brain is in the business of discovering information about the likely causes of sensory signals, to which it does not have direct access, in order to support adaptive behavior. To do this, predictive processing theories propose that the brain engages in probabilistic inference on the causes of sensory signals which in turn induces a generative model of the data via the minimization of free energy (Friston, 2009).

The viewpoint sketched above is a traditional cognitivist interpretation, but there are a variety of interpretations of predictive processing and the free energy principle, ranging from fully cognitivist (i.e. representationalist and computationalist) to radical embodied and enactive theories that do not include representation (Allen & Friston, 2018; Goldman, 2012). The literature on embodied and situated cognition, including work on SIF, has worked on interpreting and accommodating the free energy principle (Allen & Friston, 2018; Bruineberg, Kiverstein, et al., 2018; Bruineberg, Rietveld, et al., 2018; Bruineberg & Rietveld, 2014; Kirchhoff, 2018; Seth, 2013). This work argues that it is not that knowledge of the world is represented in a generative model housed in our heads, but instead claims that the full body-brain system itself constitutes a generative model of its ecological niche (Bruineberg, Kiverstein, et al., 2018; Bruineberg, Rietveld, et al., 2018; Bruineberg & Rietveld, 2014; Kirchhoff, 2018). This follows Friston (2013) when he says that "an agent does not have a model of its world—it is a model." (p. 213). This seems to be a more justified view than to think that the world is modeled exclusively inside the skull. Over time, evolutionary pressures have selected for particular features in organisms such that they can respond adaptively to their unfolding environment (Bruineberg & Rietveld, 2019a). For this to happen the whole organism (not just the brain) is shaped by the structure of the environment.

The traditional computationalist view holds that the structure shaping the generative model is the causal-probabilistic structure of the environment, i.e. the structure of the hidden causes of our sensory inputs (Kiefer & Hohwy, 2019). I don't believe this to be a productive view as it completely *lacks a concept of 'value' or 'relevance'*. We need such notions in order to determine what causal-probabilistic structure of the environment is modeled, because, surely, we do not model all causal-probabilistic structures in the environment. Why would our generative models, for instance, model the individual interactions between water molecules[35]? There seems no reason for evolutionary pressures to lead to that. Instead, what needs to be modelled are those causal-probabilistic structures that are relevant to the capacities for action an organism has, that is, they need to be relevant to affordances.

Now we can reinterpret the view of computational psychology that the generative model mimics or captures the causal-probabilistic structure *of the environment*. When I interpret causality from the viewpoint of the agent (as with my affordance-based account in Chapter 7), causality denotes pathways in the environment that allow for effective interventions for that agent. There are of course exceptions (see Section 8.3), but generally when we identify causes in daily life, we are identifying particular opportunities for action (even if we do not engage with them at the time). Using this view of what causality is for the mind allows us to see that the causal-

[35] Of course you can learn about the causal interactions between water molecules, probably by learning from the practices of physics and chemistry, and in those practices there are the skills and tools available to intervene upon those interactions. The point is that, generally, in daily life there is no reason to engage with those causal relationships as they do not relate to relevant affordances and so they would not be modeled by the agent.

probabilistic structure is not some feature of the environment, but instead a feature of the *animal-environment system*. The structure that is relevant in the environment is *structure insofar as it pertains to effective interventions* for an agent. This is what a generative model should be about, as it includes a notion of 'value' that the computationalist framework is missing.

This view is quite different from the one underlying traditional psychology, which raises the question: Is the use of cognitive models involving specific symbolic representations, such as the BMS in Chapters 4 and 5, commensurable with this view of generative models? I believe so, though it requires a pragmatist (i.e. anti-realist) view on psychological representation. Simply put, I believe cognitive models (such as the BMS and other CBN-derived models) are useful as a tool for understanding the mind as they approximate some feature of the mind. 'Some feature of the mind' is rather vague, so allow me to be a bit more speculative in order to be more concrete.

The BMS stipulates that we generate states of a causal structure which represent causal relationships in the world. Such causal relationships are traditionally interpreted as statistical features of the external world (as I do in Chapters 4 and 5), but we we are now in the position to re-interpret that. On the ecological interpretation of the free energy principle, the generative model is a model of the ecological niche in terms of relevant fields of affordances. In addition, my affordance-based account of causal engagement argues for the psychological reality of causation to be related to affordances; causation allows us to intervene on aspects of the environment effectively. This allows us to interpret the instances of 'causal structure' that are sampled not as an objective structure of the external environment, but as a structure of affordances (which relate abilities of the agent to the environment).

According to the original formulation of the BMS we generate samples via a generative model and that these samples are concrete causal system states. My tentative proposal is that the BMS works well as a model since these concrete causal system states approximate the structure of the world insofar as it relates to our action possibilities. Hence, what the model might approximate is a form of sampling or dynamic changes over anticipated fields of affordances, that is, over possible states of the world in terms how we can act on them. This could be a mechanism underlying embodied anticipation of the future (van Dijk & Rietveld, 2021a) that allows us to adaptively respond to future states of the world. Doing so does not require us to represent those states of the world directly, it requires us to anticipate our own future actions that could improve our situation (which themselves are obviously related to future states of the world).

This idea of implementing anticipation by sampling fields of affordances needs to be fleshed out much further before it can be considered a (consistent) theory. There are likely many implications and conceptual issues that I left out of the picture I sketched. Developing this theory and fleshing it out seems like a major endeavor, but it seems promising as it would possibly allow for connecting large bodies of research in cognitive psychology and embodied cognition.

If we accept the above view, we not only obtain footing for possible theoretical advances, but it can also help guide empirical research on causal learning in naturalistic settings. To see how my view can guide empirical research we first need to understand how typical experiments differ from real life. In a typical experiment on causal structure learning participants are first provided with a set of (candidate) causal variables and subsequently with information that allows participants to induce in some way what the causal relationships are between those variables (e.g. Bramley, Gerstenberg, Mayrhofer, et al., 2018; Davis et al., 2020; Griffiths & Tenenbaum, 2005; Rottman, 2017). This differs from causal cognition outside the experimental laboratory in that we normally are not given a set of variables to consider. Instead, we are confronted with the world at large and need to establish for ourselves what the relevant aspects of the environment are. This relates to the famous *causal selection problem* (Hesslow, 1988), which concerns what we should pick out as causes for an event even though every event has an infinite amount of causes if we treat causality as an objective feature of the environment (the amount of possible causes is infinite because we can trace back a causal chain as far back as the big bang for any event and then pick any event on this chain as a cause of the event). For example, if we want to pick a cause for me moving my hand, we can pick my intention (i.e. because I wanted to), but we would also be justified to pick the firing of a particular neuron, the development of my arm over my childhood, or the fact that I was born. All these options can be correct if we treat causality as an objective feature of the world, because, put simply, without them I would have not moved my arm. This makes it hard for cognitive psychology to model how people learn the causal structure of a situation in daily life simply because there exist so many aspects of the environment that can be considered.

We can constrain this set of variables (i.e. aspects of the environment) to consider and still use existing cognitive models if we accept my proposal that the (graphical) representations used in those models approximate the affordance structure in the environment for a particular agent. An agent parses the environment in such a way that it allows for adaptive action in a specific situation and we could aim to mimic that parsing by being selective in what aspects of the environment we include in a model. Doing so allows us to restrict the set of relevant environmental features to only include those features that can specify relevant affordances. For example, if we want to use a CBN model to understand the psychology behind a real life situation in which someone tries to fix a broken printer, we now have some guidance on what to include in the CBN model. We know to only include nodes which describe aspects of the environment that the agent can act upon, directly or indirectly, in addition to a node for the relevant outcome variable (i.e. whether the printer works). For instance, we would not include a node with the fact that the printer is rectangular, as the agent is unlikely to be able to act on that, but we could include a node referring to whether there is ink in the ink reservoir, as this can be manipulated by the agent (by filling the ink). This CBN model then implements the fact that the agent is likely to engage in a reasoning process involving the state of the ink reservoir, but not involving the overall shape of the printer. Such a CBN representation could then form the basis for sampling as proposed by the BMS.

More generally, using affordances to guide what to include in a CBN (or other) model allows us to fill in the model from two directions. We have a set of environmental features that we can act upon, and we have one or more relevant outcome variables. The trick for modelers and reasoners alike is then to find how those can be linked. This will involve pinning down how someone’s attention is guided through the environment to make those connections. Doing so can be a complex affair as such an account needs to incorporate the skills and practices of the agent. I do not have a concrete mechanistic suggestion for how this all takes place, but I do suggest that we use causal cognition exactly for this challenge in everyday life; the challenge of linking our action possibilities to relevant outcomes.

Lastly, one thing I hope this discussion brings home is that we should not relegate conceptual frameworks to the wastebin if we perceive them to be inconsistent with other frameworks we favor or are simply more familiar with. While an entirely consistent view that can explain all aspects of the mind is desirable, we are far from developing such a view. In the meantime, we need to accept that multiple inconsistent frameworks can each provide us with insights into the mind. For now, such insight is more valuable than consistency.

Appendices

Appendix A: Additional analyses for experiment 2 in Chapter 2

We conducted the same regression analyses used for Experiment 1 for Experiment 2 as well. In the main text only the main results of these analyses relating to the reasoning errors are presented. The other results for Experiment 2 are concisely presented here, following the same structure as for Experiment 1 in the main text. For more information about the regression analyses see Experiment 1 in the main text.

Manipulation check

To test whether the time pressure manipulation impacted response times we regressed the Deadline factor on RTs, and we found that the effect of Deadline is significant ($F(2, 3999) = 247$, $p < .001$, $\mathrm{BF}_{10} > 100$).

Overall SAT

Next we investigated the overall SAT, that is, the influence of RTs and time pressure on overall accuracy. We found a significant main effect of Deadline ($\chi^2(2) = 11.3$, $p = .004$), indicating a macro-SAT. Participants were more accurate when there was less time pressure. Post-hoc contrasts revealed that this is due to participants being significantly more accurate in the DL20 condition ($M = 13.2$, $SE = 0.924$) than in the DL6 condition ($M = 14.7$, $SE = 1.04$, $z_{\mathrm{DL6\text{-}DL20}} = 3.35$, $p = 0.002$). Accuracy in the DL9 condition does not significantly differ from the other conditions ($M = 13.8$, $SE = 0.968$, $z_{\mathrm{DL6\text{-}DL9}} = 1.84$, $p = 0.159$, $z_{\mathrm{DL9\text{-}DL20}} = 1.70$, $p = .204$). There was no significant interaction effect of RT and Deadline ($\chi^2(2) = 1.99$, $p = .369$), and the main effect of RT was just not significant ($\chi^2(1) = 3.60$, $p = .058$).

SAT Markov independence and explaining away

Markov violations Common cause and Chain

We found a significant main effect of ScreenedOff ($F(2, 1171) = 84.0$, $p < .001$), indicating that participants did not screen off, and thus violated Markov independence. The interactions of ScreenedOff with Deadline ($F(4, 1171) = .959$, $p = .429$, $\mathrm{BF}_{01} = 48.8$) and RT ($F(2, 1177) = 2.73$, $p = .065$, $\mathrm{BF}_{01} = 3.13$) were both not significant, indicating that the violations of Markov independence were not impacted by time pressure nor response times. We did find a significant interaction between ScreenedOff and MidVar ($F(2, 1171) = 72.2$, $p < .001$, $\mathrm{BF}_{10} > 100$)), indicating that the violations of Markov dependence were larger when the middle variable was present than it was not.

Markov violations Common effect

We again only found a significant main effect of ScreenedOff (*F*(2, 262) = 8.69, *p* < .001, BF_{10} > 100), indicating that participants violated Markov independence here. The interactions with ScreenedOff were not significant for both Deadline (*F*(4, 262) = 1.08, *p* = .368, BF_{01} = 9.93) and RT (*F*(2, 265) = 1.73, *p* = .180, BF_{01} = 3.89), indicating that there are no time pressure effects.

Failures to explain away

We found a significant main effect of AwayVar (*F*(2, 271) = 498, *p* < .001, BF_{10} > 100), indicating that participants did not engage in the normative explaining away pattern. The effect of knowing that other cause was absent is -4.88% (*SE* = 3.30), which is far from the CBN prediction, which says that the probability should increase by 28.6% compared to when the state of the other cause is unknown. The effect of knowing that it is present is +6.22% (*SE* = 3.38), which again is far from the CBN prediction of -11.4%.

There was no influence of deadlines on how participants explained away (*F*(4, 271) = 1.18, *p* = .318, BF_{01} = 11.7). However, we did find some evidence of an interaction of AwayVar with RT (*F*(2, 276) = 8.22, *p* < .001, BF_{10} = 0.986), as we found in Experiment 1 (see results in main text).

SAT conservative inferences

Participants tended to respond conservatively, moving on average 5.0% (*SE* = 1.10, *t* = 4.56, *p* < 0.001) towards 50% from the normative response.

We found mixed evidence of an interaction of Deadline and RTs on conservative responding (*F*(2,1676) = 4.55, *p* = .011, BF_{10} = 0.739). Focusing on main effects, we find that there is no effect of Deadline on conservatism (*F*(2,1673) = 1.93, *p* = .15, BF_{01} = 21.3), but we find a large effect of RT (*F*(1,1681) = 21.5, *p* < .001, BF_{10} > 100) indicating that conservatism is sensitive to internal time pressure. Using post-hoc contrasts, we found that the effect of RT is significant in the 6s (β = 2.18, *SE* = 0.622, *t*(1678) = 3.51, *p* < .001) and 9s deadlines (β = 1.80, *SE* = 0.454, *t*(1675) = 3.96, *p* < 0.001), but not for the 20s deadline (β = 0.549, *SE* = 0.293, *t*(1678) = 1.84, *p* = .066). Pairwise contrasts revealed that the effects in the 6s and 9s conditions are not significantly different (*t*(1675) = 0.502, *p* = . 87), while they were different from the 20s condition (versus 6s: *t*(1678) = 2.40, *p* = .044; versus 9s: *t*(1675) = 2.35, *p* = .049). Hence there seemed to be a micro-SAT for conservative inferences in the 6s and 9s conditions, but not in the 20s condition.

APPENDIX B: PARAMETER RECOVERY STUDY FOR THE BMS IN CHAPTER 4

To assess whether the Bayesian Mutation Sampler (BMS) is identifiable and what the best way of fitting it to data is we conducted a parameter recovery study. We simulated data using the BMS and then fitted the BMS to the simulated data to test whether the fitted parameters are similar to those used for simulating the data.

We test two methods of fitting the BMS to data: (1) a method employing a traditional iterative optimization approach, and (2) a method that uses a two-step grid search. Both these methods make use of the PDA method to compute 'synthetic' likelihoods (See main text; Holmes, 2015; Turner & Sederberg, 2014). Noteworthy is that each of these methods make use of the fact that the BMS has a strongly restricted parameter space, i.e. it has only two free parameters and one of those (the chain length) is an integer.

To assess the extent to which the fitting methods recover the simulated parameters we compute correlations between the true and fitted parameters. We will deem correlations below .5 to be poor, between .5 and .75 to be fair, between .75 and .9 to be good, and above .9 to be excellent, similar to criteria used in other parameter recovery studies (e.g. Anders et al., 2016; van Maanen et al., 2021; van Ravenzwaaij & Oberauer, 2009; White et al., 2015).

Simulating data

To simulate data we picked chain length parameters from a range of 2 to 50 and β parameters from a range of 0 to 15. We randomly sampled 50 values for each of the parameters from a uniform distribution over their ranges and randomly paired these. In this way we obtained 50 unique combinations of the chain length and Beta parameters which were used to simulate data.

Datasets were simulated using the causal parameters (i.e. base rates and causal strengths) of the experimental study that we fitted the BMS to in the main text (Kolvoort, Fisher, et al., 2023; these are the same causal parameters as used in Experiment 1 by Rottman & Hastie, 2016). That is, we simulated data separately for the for the Common Cause/Chain network (these had equivalent parametrizations) and for the Common Effect network.

Next, we simulated datasets with 27 and 54 observations per parameter combination, reflecting either 1 or 2 observations per inference per participant (there are 27 different inferences in the experiment by Kolvoort et al.). The smaller dataset has the same number of observations per parameter combination as the empirical data that we fitted the models to in the main text has per participant. Hence if we find good recovery for these smaller datasets, we can be confident in the parameters we obtain from fitting the models to the empirical data.

In all we simulated four data sets, two for each of the causal network structures with either 27 or 54 observations per parameter combination.

Method 1: iterative optimization of Beta prior parameter for each chain length

The first method we test can be considered more traditional as it uses an iterative method to find the optimal β parameter. In a first step we optimized the β parameter for each possible chain length, that is for each integer from 2 to 50. This optimization was done using the PDA method (Holmes, 2015; Turner & Sederberg, 2014) and the base R function *optimize*, which uses an iterative method combining golden-section search and successive parabolic interpolation (Brent, 1973; R Core Team, 2019). In this way we end up with the best fitting β parameter for each possible chain length. In the next step we simply pick the chain length and optimized β parameter that maximize the summed likelihood.

Method 2: two-step grid search

We wanted to test a second method that is robust to local minima in the likelihood landscape. The reason for this is that iterative optimization procedures can get stuck in local minima and not find the globally optimal parameters. We chose to test a two-step grid search method with additional iterative optimization (cf. Mestdagh et al., 2019). Table B1 gives an overview of the method.

We first construct a coarse parameter grid, using values ranging from 2 to 50 with step size 2 for the chain length parameter, and values from 0 to 15 with step size 1 for the β parameter. This results in a grid of 16 by 25, with (25 x 16 =) 400 unique parameter combinations. In the grid we save the predictions of the BMS under each of the unique parameter combinations. To generate these predictions we simulated 10,000 responses for each parameter combination, resulting in predicted distributions of each inference)

Next, we use the PDA method (Holmes, 2015; Turner & Sederberg, 2014) to compute the likelihood of the data for each of the parameter combinations in the coarse grid, which provides us with the best fitting ‘coarse’ parameters. Since we check each of the possible parameter combinations this method can be seen as a ‘brute force’ method. That the BMS has only 2 free parameters, of which one is an integer with a restricted range, allows for the use of such a method.

After finding the best-fitting parameter combination in the coarse grid, we construct a fine grid around this best fitting point. This fine grid consists of 7 chain lengths, the optimal coarse plus or minus 3, and of 11 β values, the optimal coarse one plus or minus 5 with a step size of 0.2. We end up with a fine grid of (7 x 11 =) 77 parameter combinations centered at the optimal parameters in the coarse grid.

As in the case of the coarse grid, we compute the likelihood for each of the 77 parameter combinations in the fine grid using the PDA method. We pick the parameter combination with the highest likelihood to obtain the best fitting parameters in the fine grid.

As a last step, we optimize the β parameter, constrained between the optimal fine grid parameter plus or minus 1, to obtain a more fine-grained estimate of β. This optimization is done iteratively as in Method 1. This last step is not done for the chain length as it is an integer.

	Step	Description
1a	Make coarse grid	Create 25 by 16 parameter grid with BMS predictions under 400 unique chain length and beta parameter combinations.
1b	Fit to coarse grid	Select optimal parameter combination from the coarse grid by maximizing the summed likelihood using the PDA method.
2a	Make fine grid	Create a 7 by 11 fine grid centered on the optimal parameter combination in the coarse grid.
2b	Fit to fine grid	Select optimal parameter combination from the fine grid by maximizing the summed likelihood using the PDA method.
3	Optimize β	Iteratively optimize β parameter, restricted to range plus or minus the optimal fine grid β parameter, using the chain length found in step 2b.

Table B1 *Overview of Method 2: the two-step grid search method*

Results Method 1

Table B2 presents the correlation coefficients between fitted and true parameters for each of the four datasets for Method 1. Figure B1 presents scatterplots of the fitted and true parameters for the datasets with 27 observations per participant.

For each of the four datasets we find poor correlations (below .423), for the chain length parameter, and poor to fair correlations for the β parameter (between .333 and .607). Together these findings indicate that Method 1, a traditional iterative method, does not satisfactorily recover the true parameters that generated the data.

Causal structure	Nr. of observations	β parameter	Chain length parameter
Common cause and Chain	27	.607	.195
	54	.446	.423
Common effect	27	.526	.072
	54	.333	.368

Table B2 Pearson correlations of simulated and fitted parameters Method 1

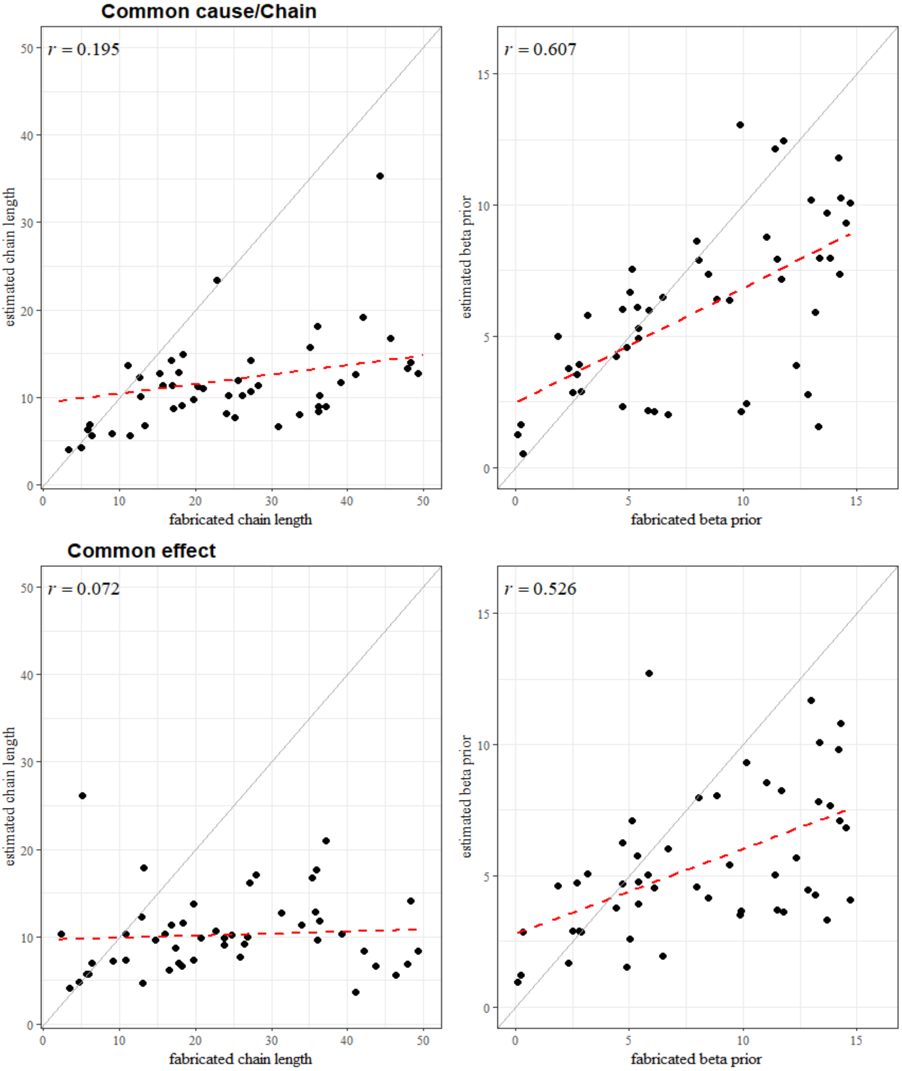


Figure B1 *Scatter plots of true and fitted parameters for Method 1 using 27 observations per participant. Grey diagonal indicates perfect recovery. Dashed red lines indicate the linear trend.*

Results Method 2

Table B3 presents the correlation coefficients between fitted and true parameters for each of the four datasets and for each of the steps of Method 2. Note that the *Optimization* column for the chain length parameter in Table B3 is intentionally left blank as the estimate of the chain length parameter does not change in this step.

Causal structure	Nr. of observations	β parameter			Chain length parameter		
		Coarse	Fine	Optimization	Coarse	Fine	Optimization
Common cause and Chain	27	. 893	. 895	. 897	.837	.853	-
	54	.926	.930	.938	.918	.928	-
Common effect	27	.897	. 883	. 871	.763	.781	-
	54	.949	.926	.934	.890	.894	-

Table B3 *Pearson correlations of simulated and fitted parameters Method 2. Fine and Coarse refer to the fine and coarse grids as explained in the text.*

From Table B3 we can see that Method 2 does accurately recover the true parameters, with all correlations being either good or excellent (between .763 and .949). Notably it is already in the first step of the method, i.e. in the coarse grid, that the correlations are high and that the subsequent steps provide only a marginal improvement in recovery. And this is also the case for the datasets with 27 observations per participant. While the recovery improves consistently with 54 observations, with 27 observations the correlations are already in the range of good to excellent. These findings indicate that one can fit the BMS accurately to data using 27 observations and only a single (coarse) grid.

Figure B3 presents scatterplots of the true and fitted parameters in each step for the datasets with 27 observations per participant, again the optimization step is left blank for the chain length parameter. From Figure B3 one can see a notable pattern of lower chain lengths being consistently more accurately estimated than higher chain lengths.

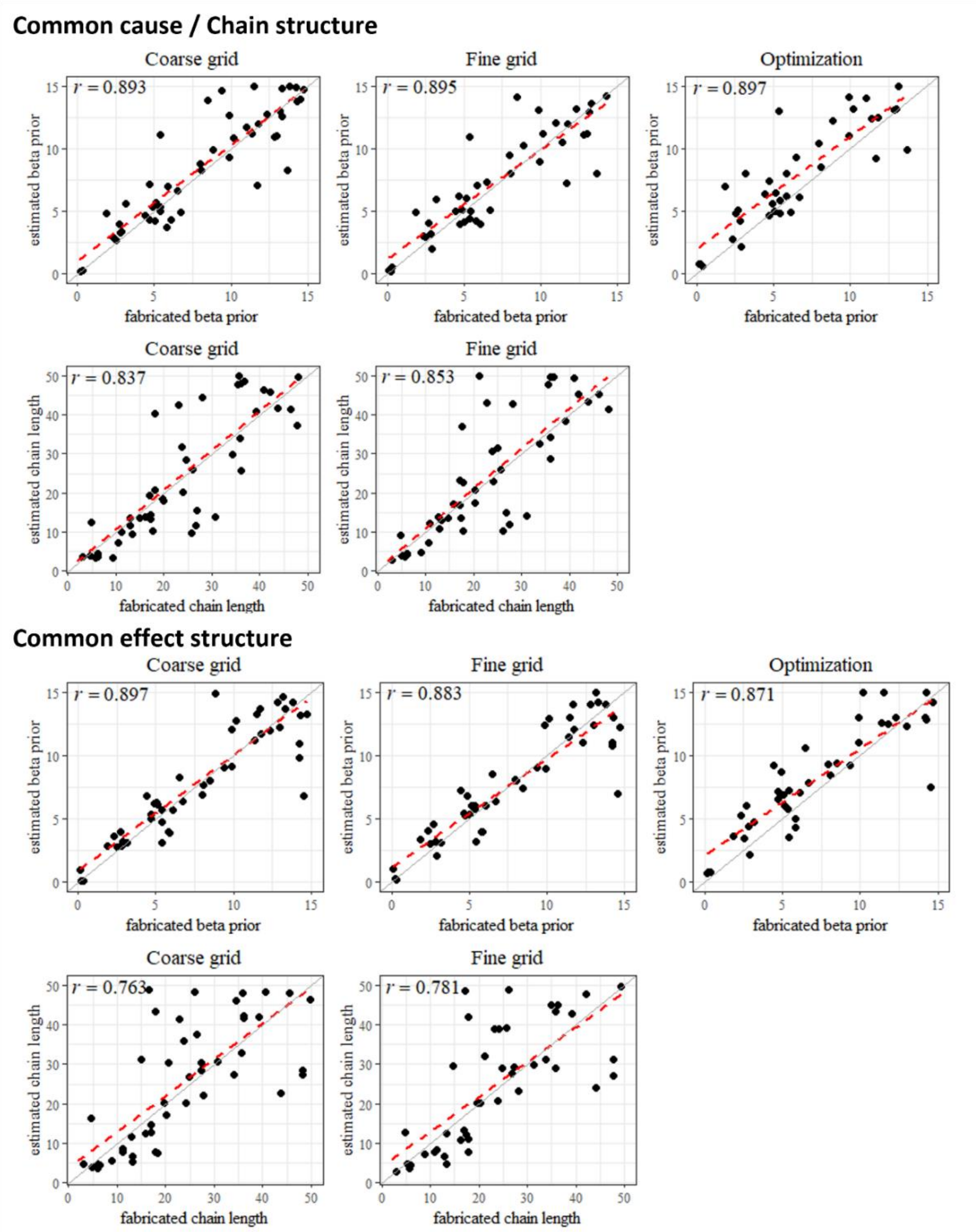


Figure B2 *Scatter plots of true and fitted parameters for Method 2 using 27 observations per participant. Grey diagonal indicates perfect recovery. Dashed red lines indicate the linear trend.*

Appendix C: Inferences per inference group in Chapter 4

Here we provide a list of inferences per inference group and their normative probability. The causal networks in the experiment were highly symmetric, allowing us to collapse over the terminal variables (e.g. $P(Y=1|X_1=1,X_2=0)=P(Y=1|X_1=0,X_2=1)$), over the presence or absence of variables (e.g. $P(Y=1|X_1=1,X_2=1)=1-P(Y=1|\,X_1=0,X_2=0)$), and over unknown variables (e.g. $P(X_1=1|\,Y=1)=P(Y=1|\,X_2=1)$). Responses to inferences with an asterisk below (*) are flipped around the midpoint to the upper portion of the probability scale based on the symmetry between the absence and presence of variables (e.g. 25% was converted to 75%; see Rottman & Hastie, 2016; Davis & Rehder 2020). Within each group all inferences have the same normative answer (after flipping) and the BMS predicts the same distribution (after flipping) for each inference in a group.

Conflict trials 1

Inferences with conflicting conditioning information where a terminal variable is queried.
Normative probability: 75%

$P(X_1=1|Y=1,\ X_2=0)$
$P(X_2=1|Y=1,\ X_1=0)$
$P(X_1=1|Y=0,\ X_2=1)$*
$P(X_1=1|Y=0,\ X_2=1)$*

Conflict trials 2

Inferences with conflicting conditioning information where the middle variable is queried.
Normative probability: 50%

$P(Y=1|X_1=1,\ X_2=0)$
$P(Y=1|X_1=0,\ X_2=1)$

Ambiguous trials 1

Inferences where the status is of only one variable is known and this variable is adjacent to the queried variable.
Normative probability: 75%

$P(X_1=1|Y=1)$
$P(X_2=1|Y=1)$
$P(X_1=1|Y=0)$*
$P(X_2=1|Y=0)$*
$P(Y=1|X_1=1)$
$P(Y=1|X_2=1)$

$P(Y = 1|X_1 = 0)$*
$P(Y = 1|X_2 = 0)$*

Ambiguous trials 2

Inferences where the status is of only one variable is known and this variable is not adjacent to the queried variable.
Normative probability: 62.5%

$P(X_1 = 1|X_2 = 1)$
$P(X_2 = 1|X_1 = 1)$
$P(X_1 = 1|X_2 = 0)$*
$P(X_2 = 1|X_1 = 0)$*

Consistent trials 1

Inferences with consistent conditioning information where a terminal variable is queried.
Normative probability: 75%

$P(X_1 = 1|Y = 1,\ X_2 = 1)$
$P(X_2 = 1|Y = 1,\ X_1 = 1)$
$P(X_1 = 1|Y = 0,\ X_2 = 0)$*
$P(X_2 = 1|Y = 0,\ X_1 = 0)$*

Consistent trials 2

Inferences with consistent conditioning information where the middle variable is queried.
Normative probability: 90%

$P(Y = 1|X_1 = 1, X_1 = 1)$
$P(Y = 1|X_1 = 0, X_1 = 0)$*

Base rates

Inferences where no conditioning information is provided.
Normative probability: 50%

$P(X_1 = 1)$
$P(X_2 = 1)$
$P(Y = 1)$

Appendix D: Fit of Mutation Sampler with scaling parameter for Chapter 4

The MS was originally fitted using a free ‘scaling’ parameter s such that a predicted response = $s*p$, where p is the predicted probability by the MS (Davis & Rehder, 2020). In the main text we have fitted the MS without such a scaling parameter as the scaling parameter can result in part of the predicted distributions falling outside of the response scale. When $s > 100$ (in the original paper it is allowed to vary between 0 and 300), responses above 100% could be produced. For instance, when $s = 150$, and the probability produced by the MS is .90, the MS with scaling would predict a response at 150 x .90 = 135% as we used 0-100% response scale. One could truncate the resultant predicted distribution (i.e. remove the responses above 100% from the prediction), but there is no psychological justification to do so (nor do the original authors do this). As such, the MS with this scaling parameter cannot provide a proper account of the variability of responses as it predicts responses to be outside the response scale.

However, it would behoove us to show that the BMS outperforms the ‘published’ version of the MS, which includes a scaling parameter. To this end we fitted the MS with a scaling parameter to the data and compared its performance with the BMS.

To fit the MS with a scaling parameter we used the same procedure as in the main text. Davis and Rehder (2020) report scaling parameter estimates ranging from 98 to 130. Based on this we chose a range for s symmetric around s = 100 (a value of 100 is equivalent to using no scaling parameter) from 70 to 130. We picked 21 values (as we did for the β parameter) equally spaced in this range, leading to the following set of values for s on the grid: [70, 73, 76, 79, 82, 85, 88, 91, 94, 97, 100, 103, 106, 109, 112, 115, 118, 121, 124, 127, 130].

We find that for 78.0% of the fits the BMS has a lower BIC than the MS with scaling (mean $\Delta_{\mathrm{BIC}} = -14.2$). We computed BIC weights as approximations for posterior model probabilities for each participant (Figure D1). For 38 out of 43 participants (88.4%) the BMS has a higher posterior probability than the MS with scaling factor. Together, these results indicate that the BMS outperforms the MS with scaling.

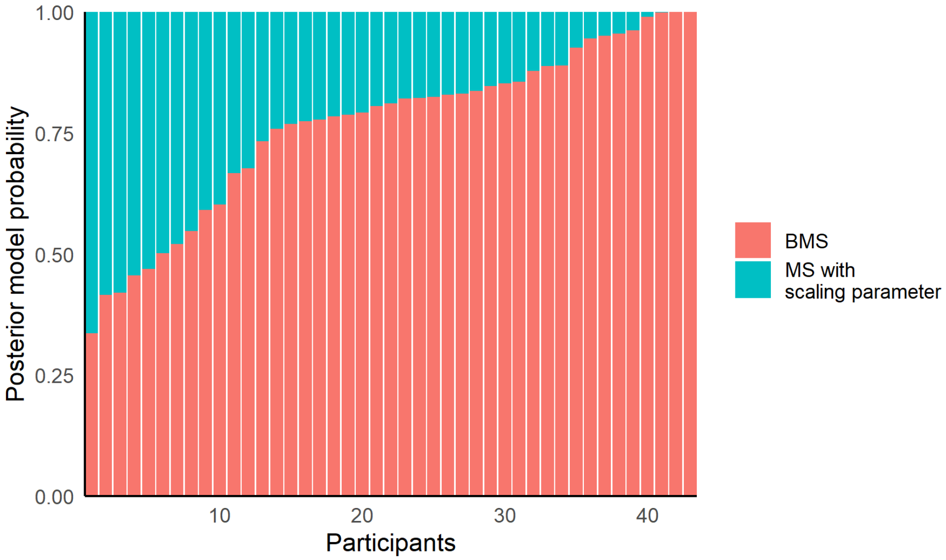


Figure D1 *Posterior model probabilities per participant comparing the BMS and the MS with scaling parameter. Posterior model probabilities are approximated using BIC weights.*

Lastly, we find that the mean fitted scaling factor is larger than 100 (*M* = 104.9, *SD* = 14.2). This indicates that indeed the model predicts responses above 100%. Regarding the five participants for whom the MS with scaling factor fit better than the BMS, we find that four of them have an average scaling factor larger than 100 (values: 113.7, 103.7, 113.0, 105.3, 88.0). For these four participants the MS with scaling factor fits better but the model predicts responses above 100%. To illustrate this, Figure D2 plots the predicted distribution for the inference $P(X_2 = 1|Y = 1,\ X_1 = 0)$ (Conflict trials 1) for one of the participants for whom the MS with scaling was the best fitting model.

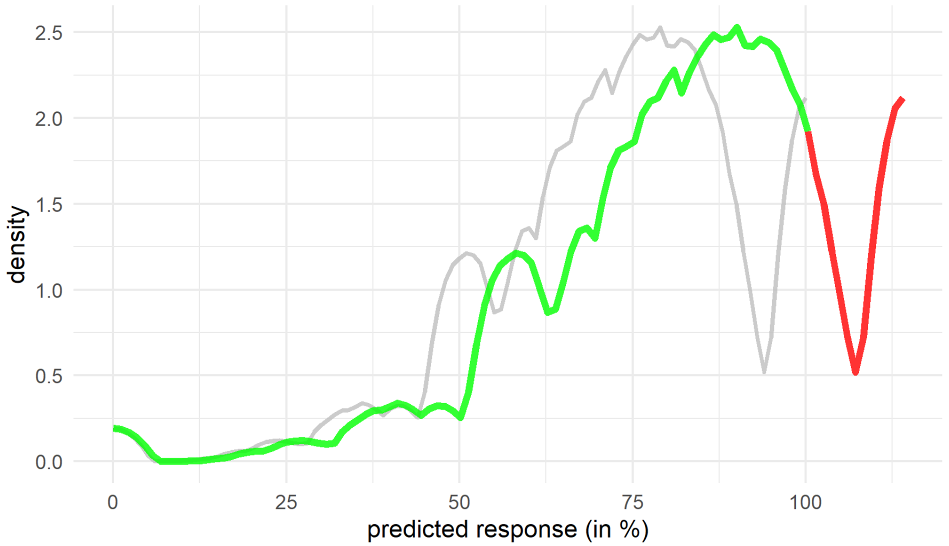


Figure D2 *Predicted distribution of MS with scaling factor. Colored line represents predicted distribution of responses of MS with scaling factor model for inference* $P(X_2 = 1|Y = 1,\ X_1 = 0)$ *using parameters: chain length = 64, s = 113. These parameters are the best fitting parameters for a participant for whom the MS with scaling factor was the best fitting model. The green part of the line indicates part of the distribution that falls within the response scale (0-100%), the red part falls outside the response scale. The thin gray line represents the predicted distribution using the same chain length but without scaling.*

AUTHOR CONTRIBUTIONS PER CHAPTER

Chapter 2
This chapter has been adapted from: Kolvoort, I.R., Fisher, E.L., Van Rooij, R.A.M., Schulz, K., & van Maanen, L. (Under review). Probabilistic Causal Reasoning under Time Pressure. Preprint DOI: 10.31234/osf.io/ej26r

Author contributions[36]
Conceptualization: IK, LVM, RVR, and SK. Methodology: IK, RVR, SK, and LVM. Software: IK, EF, and LVM. Formal analysis: IK, EF, and LVM. Investigation: EF and IK. Writing – original draft preparation: IK. Writing – review and editing: IK, RVR, SK, and LVM. Supervision: RVR, SK, and LVM.

Chapter 3
This chapter has been adapted from: Kolvoort, I.R., Davis, Z.J., van Maanen, L., & Rehder, B. (2021). Variability in Causal Judgments. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

Author contributions
Conceptualization: BR, IK, LVM, and ZD. Methodology: BR, IK, LVM, and ZD. Software: IK and ZD. Validation: BR and LVM. Formal analysis: IK and ZD. Investigation: IK and ZD. Writing – original draft preparation: IK and ZD. Writing – review and editing: BR, IK, LVM, and ZD. Supervision: BR & LVM.

Chapter 4
This chapter has been adapted from: Kolvoort, I.R., Temme, N., & van Maanen, L. (In press). The Bayesian Mutation Sampler Explains Distributions of Causal Judgments. *Open Mind*. Preprint DOI: 10.31234/osf.io/9kzb4.

Author contributions
Conceptualization: IK and LVM. Methodology: IK and LVM. Software: IK, NT, and LVM. Formal analysis: IK, NT, and LVM. Investigation: IK, NT, and LVM. Writing – original draft preparation: IK. Writing – review and editing: IK and LVM. Supervision: LVM.

Chapter 5

[36] Contributions per chapter are described using categories from the Contributor Roles Taxonomy CRediT (see https://casrai.org/credit).

This chapter has been adapted from: Kolvoort, I.R., Davis, Z.J., Rehder, B., and van Maanen, L. (Manuscript in preparation). Models of Variability in Causal Judgments.

Author contributions

Conceptualization: BR, IK, LVM, and ZD. Methodology: BR, IK, LVM, and ZD. Software: IK, ZD and LVM. Formal analysis: IK. Investigation: IK and LVM. Writing – original draft preparation: IK. Writing – review and editing: IK and LVM. Supervision: LVM.

Chapter 6

This chapter has been adapted from: Kolvoort, I.R. & Rietveld, E. (2022). Affordances for Situating the Embodied Mind in Sociocultural Practice. In Z. Djebbara (Ed.), *Affordances in Everyday Life* (pp. 13-22). Springer, Cham. DOI: 10.1007/978-3-031-08629-8_2

Author contributions

Conceptualization: ER. Methodology: IK and ER. Investigation: IK and ER. Writing – original draft preparation: IK. Writing – review and editing: IK and ER. Supervision: ER.

Chapter 7

This chapter has been adapted from: Kolvoort, I.R., Schulz & Rietveld, E. (In press). The Causal Mind: An Affordance-Based Account of Causal Engagement. *Adaptive Behavior*.

Author contributions

Conceptualization: IK. Methodology: IK and ER. Investigation: IK, ER and KS. Writing – original draft preparation: IK and ER. Writing – review and editing: IK, ER, and KS. Supervision: ER and KS

ENGLISH SUMMARY

Novel perspectives on the causal mind: experiments, modeling, and theory

This thesis presents research into human causal cognition using a multiple perspectives and methodologies. The main research in this thesis is structured in three parts.

Part 1: Experimental Studies on Probabilistic Causal Inference
In Part 1 I present two sets of experiments on causal reasoning where we teach participants causal network information and then ask them to solve inference problems in the form of causal probabilistic queries (e.g.: "if X causes A and B, what is the probability of A being present knowing that X is but B is not present?").

Chapter 2 focusses on the effect of time pressure on such causal judgements and the errors people make reasoning this way. We find that one type of error, Markov violations, is affected by time pressure, while another, over-conservatism, is not. This was surprising, as existing theories would predict Markov violations to increase. The findings in this chapter indicate that causal inferences (and errors therein) are not the result of a single cognitive mechanism. Instead, the underlying processes are likely to be more complex.

The experiment in Chapter 3 uses multiple techniques to elicit repeated judgments for participants in order to assess, the variability in causal judgments. The results, for the first time, showed that the variability previously observed in causal judgments is due to both between- and within-participant variability. Moreover, we find that the within-participant variability is affected by the type of inference presented to participants. The important implication from this finding is that variability in causal judgments, at least partly, reflects the judgment process itself rather than just noise. This means that the tradition to focus only on averaged behavior has led researchers away from valuable information. Instead, we should take into account (aspects of) distributions of responses when developing cognitive models of causal reasoning.

Part 2: Computational Cognitive Modeling of Causal Reasoning
In the second part of this thesis, I develop and test a new cognitive model of causal reasoning, named the Bayesian Mutation Sampler (BMS), and compare it to other models that can account for variability in causal judgments.

I start Chapter 4 by scrutinizing a recent model of causal reasoning, the Mutation Sampler (MS; Davis & Rehder, 2020). My analysis identifies that, while the MS performs well at predicting mean judgments, it fails to account for salient features of distributions of causal judgments, such as a lack of extreme responses (i.e. responses near 0% and 100%). I argue that the MS lacks a mechanism for incorporating prior information and that this is a likely reason for the misfits. I develop a generalization of the MS, the BMS, which combines the sampling procedure of the MS with the use prior information about good responses. I then test the MS and BMS on the experimental data from Chapter 2. I find that the BMS clearly outperforms the MS, in terms of predicting mean judgments as well as distributions of judgments. As it stands, the

BMS is the first model that is able to account for response distributions on probabilistic causal reasoning tasks.. These results suggest that the variability observed in causal judgments is due to the stochastic sampling scheme underlying the BMS, something I test further in Chapter 5.

In Chapter 5 I test the BMS against other candidate models to see whether they can account for the variability in causal judgements and other well-known patterns in the causal judgment data from Chapter 3. In addition to the BMS, I tested the existing Beta Inference Model and four other models I develop based on general psychological mechanisms that could produce variability in causal judgments. While there exist many other theories of causal reasoning, the ones I test seem the only ones that can produce variable judgments as we observe them in experiments. I find that, overall, the BMS outperforms all other models. Both in terms of quantitative fit and in terms of accounting for qualitative patterns of interest the BMS fares best. None of the tested models however, accounted for the changes in within-participant over the different inference types that were identified in Chapter 3. The findings suggest that the fit of the BMS can be improved by letting the amount of samples a reasoner takes vary based on the inference type. But even before that work is done, the BMS seems to already provides the best process-level account of causal reasoning in the literature.

Part 3: Affordances and Causal Engagement

In Part 3 I move away from cognitive psychology and delve into philosophy. I identify a lack of an embodied perspective on causal cognition and subsequently put forward an affordance-based theory of causal cognition rooted in the ideas from ecological psychology and enactivism.

In Chapter 6 I present a general introduction to the Skilled Intentionality Framework (SIF) and the role affordances, i.e. possibilities for action, plays therein. SIF combines the ideas from ecological psychology and enactivism to understand the situated and embodied mind. A crucial part of this is to construe affordances as relations between the abilities available to an organism and their environment. This allows the application of an affordance-based analysis to any type of skilled behavior. In this chapter I discuss how such affordance-based analyses can apply to the level of our ecological niche, at the level of a sociocultural practice, and at the individual level.

In Chapter 7 I develop and present an affordance-based account of causal engagement emphasizing its embodied and situated nature. Causal engagement, as I conceive of it, underlies most of causal judgments and perceptions as they occur in daily life. At the core of my account is that causal engagement is a skill and this skill is about selectively attending to aspects in our environment that allow for effective interventions. These effective interventions are understood as relevant affordances. Which actions are effective interventions (or: which affordances are relevant) depends on the material and sociocultural environment. Construing causal engagement this way allows us to understand the variation in causal judgments between different cultures and between people part of different practices, as being due to differences in skills, practices, and culture. Interventions that are used in one practice might not be in another, and so people inhabiting the former might experience causality in different aspects of the environment.

This affordance-based account of causal engagement provides a theoretical framework for understanding how we experience causation and it has a broader scope than Chapters 2-5. Chapters 2-5 used the traditional conceptual framework of cognitive psychology, which conceives of causal cognition primarily in terms of processing of (statistical) information. However, solely using an information processing metaphor to understand the mind prohibits grasping the embodied, situated and, enacted nature of how we deal with causality in daily life. My affordance-based account is not at odds with the conceptual framework used in Chapters 2-5, but encompasses it and describes causality and its role in cognition at a more fundamental level. The view used in Chapters 2-5 focuses on our immensely impactful capacity to use statistical information and judge probabilities in the context of causal structures. However, there is more to causal cognition (and the mind) than that. We do not have a consistent theory that explains everything about the mind, until we do we should accept that different, possibly inconsistent, frameworks can provide valuable insight into the mind.

NEDERLANDSE SAMENVATTING

Nieuwe perspectieven op de causale geest: experimenten, modellen, en theorie

In dit proefschrift presenteer ik onderzoek naar menselijke causale cognitie vanuit verschillende perspectieven en gebruik makend van verschillende methodologieën. Causale cognitie verwijst naar het vermogen van mensen om oorzaak-en-gevolgrelaties te begrijpen en te redeneren over hoe gebeurtenissen met elkaar verbonden zijn in termen van oorzaken en effecten. Dit omvat het vermogen om causaliteit te herkennen, te voorspellen en te verklaren. Dit is essentieel voor ons denken, het maken van beslissingen, en het oplossen van problemen, zowel in het dagelijks leven als in de wetenschap. Als mensen zijn we continu bezig om de structuur van oorzaken en gevolgen om ons heen te begrijpen. Of we nou een verjaardagsfeest willen organiseren, een lekker broodje willen maken, of een raket de ruimte in willen schieten, het begrijpen van de relevante oorzaken (dat is, de oorzaken van een succesvol feest, lancering, of wat nou precies een broodje lekker maakt) is cruciaal.

Het onderzoek in dit proefschrift is gestructureerd in drie delen, elk bestaande uit twee hoofdstukken.

Deel 1: Experimentele studies naar Probabilistische Causale Gevolgtrekking

In Deel 1 presenteer ik twee reeksen experimenten over causaal redeneren waarbij ik proefpersonen causale netwerkinformatie aanleerde en hen vervolgens vroeg inferentieproblemen op te lossen in de vorm van causale probabilistische vragen zoals "als X de oorzaak is van A en B, wat is dan de kans dat A aanwezig is wetende dat X wel aanwezig is maar B niet?".

Hoofdstuk 2 richt zich op het effect van tijdsdruk op zulke causale oordelen en de fouten die mensen maken als ze causaal redeneren. Ik vond dat één type fout, Markov-overtredingen, beïnvloed werd door tijdsdruk, terwijl een ander type fout, over-conservatisme, niet beïnvloed werd. Dit was verrassend, omdat reeds bestaande theorieën voorspellen dat Markov-overtredingen zouden moeten toenemen onder tijdsdruk. De bevindingen in dit hoofdstuk geven aan dat causale gevolgtrekkingen (en fouten daarin) niet het resultaat zijn van één enkel cognitief mechanisme. In plaats daarvan zijn de onderliggende processen een stuk complexer.

Het experimentele werk in Hoofdstuk 3 gebruikte speciaal ontwikkelde technieken om meerdere, identieke oordelen te ontlokken van proefpersonen om de variabiliteit in causale oordelen te analyseren. De resultaten lieten voor het eerst zien dat de eerder waargenomen variabiliteit in causale oordelen het gevolg is van zowel variabiliteit *tussen* proefpersonen als variabiliteit *binnen* proefpersonen. Verder vond ik dat de variabiliteit binnen de proefpersonen wordt beïnvloed door het type gevolgtrekking dat aan hen wordt voorgelegd. De belangrijke implicatie van deze bevinding is dat variabiliteit in causale oordelen, tenminste gedeeltelijk, het beoordelingsproces zelf weerspiegelt en niet alleen ruis. Dit betekent dat de traditie binnen psychologie om vooral te focussen op gemiddelde antwoorden onderzoekers heeft weggeleid van waardevolle informatie. In plaats daarvan moeten we rekening houden met (aspecten van) verdelingen van oordelen bij het ontwikkelen van cognitieve modellen van causaal redeneren.

Deel 2: Computationele cognitieve modellering van causaal redeneren

In het tweede deel van dit proefschrift ontwikkelde en testte ik een nieuw wiskundig cognitief model van causaal redeneren, genaamd de Bayesian Mutation Sampler (BMS), en vergeleek ik het met andere modellen die variabiliteit in causale oordelen mogelijk konden verklaren.

Hoofdstuk 4 begint met het onder de loep nemen van een recent model van causaal redeneren, de Mutation Sampler (MS; Davis & Rehder, 2020). Mijn analyse liet zien dat de MS weliswaar goed presteert in het voorspellen van gemiddelde antwoorden voor de inferentieproblemen, maar problemen heeft met het voorspellen van opvallende kenmerken van verdelingen van deze causale oordelen, zoals een gebrek aan extreme antwoorden (d.w.z. antwoorden in de buurt van 0% en 100%). Ik beargumenteer dat de MS een mechanisme mist voor het incorporeren van reeds bekende informatie (de zogenaamde “prior” in Bayesiaanse termen) en dat dit waarschijnlijke de reden is dat de MS de verdeling van antwoorden verkeerd schat. Om deze verklaring te formaliseren ontwikkelde ik een generalisatie van de MS, de BMS, die de sampling procedure van de MS combineert met het gebruik van een prior over wat mogelijk goede antwoorden zijn. Vervolgens testte ik de MS en BMS op de experimentele data uit Hoofdstuk 2. Ik vond dat de BMS het duidelijk beter deed dan de MS, zowel in het voorspellen van gemiddelde oordelen als in het voorspellen van verdelingen van oordelen. Op dit moment is de BMS het eerste model dat in staat is om antwoordverdelingen op probabilistische causale redeneertaken te verklaren. Deze resultaten suggereren dat de waargenomen variabiliteit in causale oordelen het gevolg is van het stochastische sampling mechanisme dat ten grondslag ligt aan de BMS, iets wat ik in Hoofdstuk 5 verder testte.

In Hoofdstuk 5 toetste ik de BMS en andere kandidaat modellen om te zien of zij de variabiliteit in causale oordelen (en andere bekende patronen) in de experimentele data uit Hoofdstuk 3 kunnen verklaren. Naast de BMS testte ik het bestaande Beta Inference Model en vier andere modellen die ik ontwikkelde op basis van algemene psychologische mechanismen die variabiliteit in causale oordelen zouden kunnen veroorzaken. Hoewel er veel andere theorieën over causaal redeneren bestaan, lijken de modellen die ik testte de enigen te zijn die variabele oordelen kunnen produceren zoals we die in experimenten waarnemen. Ik vond dat de BMS over het algemeen beter presteert dan alle andere modellen. Zowel in termen van kwantitatieve fit als in termen van het verklaren van kwalitatieve patronen van belang scoorde de BMS het beste. Op basis van deze resultaten lijkt de BMS de beste procesmatige beschrijving van causaal redeneren te zijn die de cognitieve wetenschap momenteel te bieden heeft.

Deel 3: Affordances en causale interactie

In Deel 3 bewoog ik weg van de cognitieve psychologie en dook ik de filosofie in. Hier identificeerde ik een gebrek aan een belichaamd perspectief op causale cognitie en stelde vervolgens een op affordances gebaseerde theorie van causale cognitie voor die geworteld is in ideeën uit de ecologische psychologie en het enactivisme.

In Hoofdstuk 6 presenteerde ik een algemene inleiding tot het Skilled Intentionality Framework (SIF) en de rol die affordances, d.w.z. handelingsmogelijkheden, daarin spelen. SIF combineert de ideeën uit de ecologische psychologie en het enactivisme om de gesitueerde en belichaamde geest te begrijpen. Een cruciaal onderdeel hiervan is om affordances op te vatten als relaties tussen de handelingsmogelijkheden van een organisme en zijn omgeving. Dit maakt de toepassing van een op affordances gebaseerde analyse op elk type gedrag mogelijk. In dit hoofdstuk besprak ik ook hoe zulke op affordances gebaseerde analyses kunnen worden toegepast op het niveau van onze ecologische niche, op het niveau van een socioculturele praktijk, en op het niveau van een individu.

In Hoofdstuk 7 ontwikkelde ik een op affordances gebaseerde beschrijving van causale interactie, waarbij ik de nadruk legde op de belichaamde en gesitueerde aard ervan. Zulke causale interacties, zoals ik het construeer, liggen ten grondslag aan het overgrote deel van causale oordelen en waarnemingen zoals die in het dagelijks leven voorkomen.

De kern van mijn beschrijving is dat causale interactie een vaardigheid is en dat deze vaardigheid bestaat uit het selectief letten op aspecten in onze omgeving die effectieve interventies mogelijk maken. Deze effectieve interventies worden opgevat als relevante affordances. Welke acties effectieve interventies zijn (of: welke affordances relevant zijn) hangt af van de materiële en sociaal-culturele omgeving. Door causale interactie op deze manier te construeren, kunnen we de variatie in causale oordelen begrijpen als zijnde het gevolg van verschillen in vaardigheden, praktijken, en cultuur. Interventies die gebruikelijk zijn in de ene culturele praktijk zijn dat misschien niet in een andere, en dus kunnen mensen, afhankelijk van de praktijken waar ze deel van uitmaken, causaliteit ervaren in verschillende aspecten van de omgeving.

Deze op affordance gebaseerde beschrijving van causale interactie biedt een theoretisch kader om te begrijpen hoe we causaliteit ervaren en heeft een bredere reikwijdte dan de Hoofdstukken 2 tot en met 5. De Hoofdstukken 2 tot en met 5 gebruikten het traditionele conceptuele kader van de cognitieve psychologie, dat causale cognitie voornamelijk opvat in termen van de verwerking van (statistische) informatie. Als je echter alleen een metafoor voor informatieverwerking gebruikt om de geest te begrijpen, kun je de belichaamde, gesitueerde, en uitgevoerde aard van onze omgang met causaliteit in het dagelijks leven niet begrijpen. Mijn op affordance gebaseerde beschrijving is niet per se in tegenspraak met het conceptuele kader dat gebruikt is in de Hoofdstukken 2-5, maar beschrijft causaliteit en de rol ervan in cognitie op een fundamentelere manier.

Het perspectief dat in de hoofdstukken 2-5 wordt gebruikt, richt zich op ons enorm invloedrijke vermogen om statistische informatie te gebruiken en waarschijnlijkheden te beoordelen in de context van causale structuren. Echter, causale cognitie (en de geest) omvat veel meer dan dat. We hebben op het moment geen consistente theorie die alles over de geest verklaart, en dus moeten we tot die tijd accepteren dat verschillende, mogelijk inconsistente, theoretische raamwerken een waardevol inzicht in onze psyche kunnen geven.

ACKNOWLEDGEMENTS

The causal pathways by which this thesis has come into existence are numerous and complex. Let me give you some of the highlights.

Firstly, none of this work would have been possible without my supervisors, causal forces from beginning to end, Leendert van Maanen and Katrin Schulz. I am grateful for the opportunity you have given me to do this PhD, for the kindness you have shown me, and for both guiding me throughout this process as well as giving me the freedom to go in whatever research direction I saw fit. Leendert, your broad intellectual curiosity has been positively contagious as has been your skepticism and down-to-earth approach to science. You have also been a great example to me in how to enjoy life and not become an overworked academic. Thank you, Katrin, for showing me new levels of conceptual and linguistic precision. Even though I often ventured into subdisciplines that are not the most familiar to you, your ability to challenge my conceptual understanding and provide fresh perspectives has had a major impact on all of my research.

I also want to thank Robert van Rooij and Han van der Maas, my interim promotors, for engaging me intellectually from the very start and for taking it upon yourself to smooth out any formal hurdles.

Scott, you contributed greatly to me being able to do this research while keeping most of my sanity. From before any of it started, as you were the one to tell me about this project, until after it will all be finished. I am grateful that we were able to go on our PhD journeys together and share so many of the highs and lows. Nothing of it would have been the same without the sublime crackhome we created. And thank you Pirri for being such a qualified support dog.

I'm also truly grateful for you, Deni, for walking into my life midway through this PhD and for creating the many beautiful moments we've had since. Your presence has thoroughly colored my path here in unexpected and unforgettable ways.

Thank you Dana, John, Jack, and Dan, for creating the most violently chill Sunday ritual imaginable. Thank you too, Raghav, mostly just for being weird. And you, Jaël, for being just wonderful.

Jij ook bedankt Floris, voor de vele afleidingen de afgelopen jaren en voor het zijn van mijn paranimf. En jullie ook Tom en Annerose, de vele dinertjes zorgden er voor dat ik toch enigszins met mijn voeten op de grond ben blijven staan ook als was dat soms met een gestrekt been in de lucht.

And I want to thank some fellow (ex-)PhD'ers at PML. Maarten, Fabian, Adam, and Johnny, thank you for dealing with my chatting in the office and for making me feel at home at PML. Jonas, I wouldn't have had as many laughs nor be as equanimous if it wasn't for our regular efforts at attaining corporal and mental enlightenment.

Erik Rietveld, I am truly thankful for your genuine interest in my philosophical ideas and your proactive efforts introducing me to and involving me in the world of (ecological and enactive) philosophy.

Also thank you, Bob Rehder and Zach Davis, for being outstanding scientists and collaborators.

And thank you to Denny Borsboom, Neil Bramley, Julia Haaf, Erik Rietveld, Rineke Verbrugge, and Willem Zuidema for taking the time to read through this dissertation and for being part of the doctoral committee.

En natuurlijk grootse dank voor mijn ouders. Pa en ma, bedankt voor de liefde en support, jullie zijn de ultieme oorzaak van dit alles!

BIBLIOGRAPHY

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The Future of Embodied Design for Mathematics Teaching and Learning. *Frontiers in Education*, 147.

Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617.

Ahn, W. Y., & Busemeyer, J. R. (2016). Challenges and promises for translating computational tools into clinical practice. *Current Opinion in Behavioral Sciences*, *11*, 1–7.

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403–418.

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482.

Anders, R., Alario, F. X., & Maanen, L. Van. (2016). The shifted wald distribution for response time data analysis. *Psychological Methods*, *21*(3), 309–327.

Araújo, D., & Davids, K. (2011). What Exactly is Acquired during Skill Acquisition? *Journal of Consciousness Studies*, *18*(3), 7–23.

Archambeau, K., Couto, J., & van Maanen, L. (2022). Non-Parametric Mixture Modeling Of Cognitive Psychological Data : A New Method To disentangle Hidden Strategies. *Behavior Research Methods*.

Atkinson, D. (2010). Extended, Embodied Cognition and Second Language Acquisition. *Applied Linguistics*, *31*(5), 599–622.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baetu, I., & Baker, A. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 153.

Balakrishnan, J. D. (1996). Testing Models of Decision Making Using Confidence Ratings in Classification. *Article in Journal of Experimental Psychology Human Perception & Performance*.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Beller, S., Bender, A., & Song, J. (2009). Weighin Up Physical Causes: Effects of Culture, Linguistic Cues and Content. *Journal of Cognition and Culture*, *9*, 347–365.

Bender, A. (2020). What Is Causal Cognition? *Frontiers in Psychology*, *11*(January), 1–6.

Bender, A., & Beller, S. (2011). Causal asymmetry across cultures: Assigning causal roles in symmetric physical settings. *Frontiers in Psychology*, *2*(SEP), 1–10.

Bender, A., & Beller, S. (2017). Agents and patients in physical settings: Linguistic cues affect the assignment of causality in German and Tongan. *Frontiers in Psychology*, *8*.

Bender, A., & Beller, S. (2019). The Cultural Fabric of Human Causal Cognition. *Perspectives on Psychological Science*, *14*(6), 922–940.

Bender, A., Beller, S., & Medin, D. L. (2017). Causal Cognition and Culture. In *Oxford Handbook of Causal Reasoning* (pp. 717–738). Oxford University Press.

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do humans produce the speed-accuracy trade-off that maximizes reward rate? *The Quarterly Journal of Experimental Psychology*, *63*(5), 863–891.

Bogacz, R., Wagenmakers, E. J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, *33*(1), 10–16.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10).

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in Causal Structure Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(12), 1880–1910.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*(May), 9–38.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative Forgetful Scholars : How People Learn Causal Structure Through Sequences of Interventions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(3).

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall.

Bruineberg, J., Chemero, A., & Rietveld, E. (2018). General ecological information supports engagement with affordances for 'higher' cognition. *Synthese*, 1–21.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417–2444.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*, 1–14.

Bruineberg, J., & Rietveld, E. (2019a). What's inside your head once you've figured out what your head's inside of. *Ecological Psychology*, *31*(3), 198–217.

Bruineberg, J., & Rietveld, E. (2019b). What's Inside Your Head Once You've Figured Out What Your Head's Inside Of. *Ecological Psychology*, *31*(3), 198–217.

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: A niche construction perspective. *Journal of Theoretical Biology*, *455*(July), 161–178.

Bruineberg, J., Seifert, L., Rietveld, E., & Kiverstein, J. (2021). Metastable attunement and real-life skilled behavior. *Synthese*, *199*, 12819–12842.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions Based on Numerically and Verbally Expressed Uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(2), 281–294.

Campbell, J. I. D., & Xue, Q. (2001). Cognitive Arithmetic Across Cultures. *Journal of Experimental Psychology: General*, *130*(2), 299–315.

Carnap, R. (1966). *An introduction to the philosophy of science*.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Springer*, *46*(1), 112–130.

Chater, N., Zhu, J. Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020). Probabilistic Biases Meet the Bayesian Brain. *Current Directions in Psychological Science*, *29*(5), 506–512.

Chemero, A. (2003). An Outline of a Theory of Affordances. *Ecological Psychology*, *15*(2), 181–195.

Chemero, A. (2009). *Radical embodied cognitive science*. MIT Press.

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review*, *104*(2), 367–405.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*(1–2), 83–120.

Choi, H., & Scholl, B. J. (2004). Effects of grouping and attention on the perception of causality. *Perception and Psychophysics*, *66*(6), 926–942.

Choi, I., Nisbett, R. E., Norenzayan, A., Choi, Nisbett, Norenzayan, Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal Attribution Across Cultures: Variation and Universality. *Pyschological Bulletin*, *125*(1), 47–63.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Coltheart, M., Rastle, K., Holloway, R., Perry, C., & Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480.

Couto, J., Van Maanen, L., & Lebreton, M. (2020). Investigating the origin and consequences of endogenous default options in repeated economic choices. *PLoS ONE*, *15*(8), e0232385.

Crump, M. J. C., Mcdonnell, J. V, & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), 57410.

Dandurand, F., Shultz, T., & Onishi, K. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. MIT Press.

Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning* (pp. 201–215). Oxford University Press.

Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences*, *25*(3), 240–251.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*(056), 1–25.

David, H. A. (1968). Gini's Mean Difference Rediscovered. *Biometrika*, *55*(3), 573–575.

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, *60*(23), 685–700.

Davis, Z. J., Bramley, N. R., & Rehder, B. (2020). Causal Structure Learning in Continuous Systems. *Frontiers in Psychology*, *11*, 287–292.

Davis, Z. J., & Rehder, B. (2017). The Causal Sampler: A Sampling Approach to Causal Representation, Reasoning and Learning. *CogSci 2017*, 1–6.

Davis, Z. J., & Rehder, B. (2020). A Process Model of Causal Reasoning. *Cognitive Science*, *44*(5), 1–41.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition*, *126*, 285–300.

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-Slots Models of Visual Working-Memory Response Times. *Psychological Review*, *120*(4), 873–902.

Dretske, F. (1989). Reasons and Causes. *Philosophical Perspectives*, *3*, 1–15.

Dreyfus, H., & Kelly, S. D. (2007). Heterophenomenology: Heavy-handed sleight-of-hand. *Phenomenology and the Cognitive Sciences*, *6*, 45–55.

DuCharme, W. M. (1970). Response bias explanation of conservative human inference. *Journal of Experimental Psychology*, *85*(1), 66–74.

Dunlosky, J., & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition*, *29*(2), 247–253.

Dutilh-Novaes, C. (2019). Logic and the psychology of reasoning. In M. Kusch (Ed.), *The Routledge handbook of Philosophy of Relativism* (pp. 445–454). Routledge.

Dutilh, G., Wagenmakers, E. J., Visser, I., & Van Der Maas, H. L. J. (2011). A Phase Transition Model for the Speed-Accuracy Trade-Off in Response Time Experiments. *Cognitive Science*, *35*(2), 211–250.

Edland, A., & Svenson, O. (1993). Judgment and Decision Making Under Time Pressure. In *Time Pressure and Stress in Human Judgment and Decision Making* (pp. 27–40). Springer US.

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

Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278.

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389.

Evans, J. S. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning under time pressure: a study of causal conditional inference. *Experimental Psychology*, *56*(2), 77–83.

Fengler, A., & Frank, M. J. (2020). Encoder-Decoder Neural Architectures for Fast Amortized Inference of Cognitive Process Models. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1859–1865.

Fernbach, P. M., & Darlow, A. (2010). Causal Conditional Reasoning and Conditional Likelihood. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, *1*, 1088–1093.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of Alternative Causes in Predictive but Not Diagnostic Reasoning. *Psychological Science*, *21*(3), 329–336.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in Predictive and Diagnostic Reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168–185.

Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(5), 1327–1343.

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

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

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Yves Von Cramon, D., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *PNAS*, *105*(45), 17538–17542.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*(1), 641–666.

Frijda, N. H. (2007). *The Laws of Emotion*. Lawrence Erlbaum Associates, Inc.

Friston, K. (2009). The free-energy principle : a rough guide to the brain ? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86).

Froese, T., & Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmatics & Cognition*, *19*(1), 1–36.

Furlan, S., Agnoli, F., & Reyna, V. F. (2016). Intuition and analytic processes in probabilistic reasoning: The role of time pressure. *Learning and Individual Differences*, *45*.

Gallagher, S. (2017). *Enactivist Interventions*. Oxford University Press.

Gallagher, S., & Zahavi, D. (2008). *The phenomenological mind: An introduction to philosophy of mind and cognitive science*. Routledge.

Gershman, S. J., & Goodman, N. D. (2014). Amortized Inference in Probabilistic Reasoning Amortized Inference in Probabilistic Reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 517–522.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936–975.

Gibson, E., & Pick, A. (2000). *An Ecological Approach to Perceptual Learning and Development*. Oxford University Press.

Gibson, E., & Rader, N. (1979). Attention: The perceiver as performer. In G. Hale & M. Lewis (Eds.), *Attention and cognitive development* (pp. 1–21). Plenum.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual Learning: Differentiation or Enrichment? *Psychological Review*, *62*(1), 32–41.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the Fast and Frugal Way : Models of Bounded Rationality. *Psychological Review*, *103*(4), 650–669.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., Spirtes, P., Choh, ·, Teng, M., Zhang, J., Glymour, C., Danks, · D, Glymour, B., Eberhardt, F., Ramsey, J., Scheines, · R, Spirtes, · P, Teng, C. M., & Zhang, J. (2010). Actual causation: A stone soup essay. *Synthese*, *175*(2), 169–192.

Goldman, A. I. (2012). A Moderate Approach to Embodied Cognitive Science. *Review of Philosophy and Psychology*, *3*(1), 71–88.

Goldsmith, R. W., & Sahlin, N. E. (1983). The role of second-order probabilities in decision making. *Advances in Psychology*, *14*, 455–467.

Golonka, S. (2015). Laws and Conventions in Language-Related Behaviors. *Ecological Psychology, 27*(3), 236–250.

Golonka, S., & Wilson, A. D. (2019). Ecological Representations. *Ecological Psychology, 31*(3), 235–253.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a Probabilistic Language of Thought. In *he Conceptual Mind: New Directions in the Study of Concepts* (pp. 623–653). MIT press.

Gopnik, A., & Schulz, L. (Eds.). (2007). *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley & Sons Ltd.

Grice, G. R., & Spiker, V. A. (1979). Speed-accuracy tradeoff in choice reaction time: Within conditions, between conditions, and between subjects. *Perception & Psychophysics*, *26*(2), 118–126.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-Based Causal Induction. *Psychological Review, 116*(4), 661–716.

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science, 16*(4), 789–802.

Gunawan, D., Hawkins, G. E., Kohn, R., Tran, M.-N., & Brown, S. D. (2021). *Time-Evolving Psychological Processes Over Repeated Decisions*.

Hagmayer, Y. (2016). Causal Bayes nets as psychological theories of causal reasoning: evidence from psychological research. *Synthese*, *193*(4), 1107–1126.

Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: causal Bayes nets as rational models of everyday causal reasoning. *Synthese*, *189*, 603–614.

Haines, N., Kvam, P. D., Irving, L., Smith, C., Beauchaine, T., Pitt, M., Ahn, W., & Turner, B. M. (2022). Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox. *PsyArXiv*.

Hájek, A. (2012). Interpretations of probability. In E N Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2019 Edition)*.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97–109.

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition*, *1333*, 611–620.

Heft, H. (1989). Affordances and the Body: An Intentional Analysis of Gibson's Ecological Approach to Visual Perception. *Journal for the Theory of Social Behaviour, 19*(1), 1–30.

Heft, H. (2001). *Ecological psychology in context: James Gibson, Roger Barker, and the legacy of William James's radical empiricism*.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*(8 JUN), 1–19.

Hemmer, P., Tauber, S., & Steyvers, M. (2015). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin and Review*, *22*(3), 614–628.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237.

Hesslow, G. (1988). The problem of causal selection. In *Contemporary science and natural explanation: Commonsense conceptions of causality*.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211–237.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*, 383–400.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. J. (2020). A Conceptual Introduction to Bayesian Model Averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

Hitchcock, C. (2017). Actual causation: What's the use. In *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *106*(11), 587–612.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*(4), 382–401.

Hoge, R. D. (1970). perceived accuracy of information processing. *Psychonomic Science*, *18*(6), 351–353.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Holmes, W. R. (2015). A practical guide to the Probability Density Approximation (PDA) with improved implementation and error characterization. *Journal of Mathematical Psychology*, *68–69*, 13–24.

Holyoak, K. J., & Cheng, P. W. (2011). Causal Learning and Inference as a Rational Process: The New Synthesis. *Annual Review of Psychology*, *62*, 135–163.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and Category-Based Inference: A Theoretical Integration With Bayesian Causal Models. *Journal of Experimental Psychology: General*, *139*(4), 702–727.

Hume, D. (1748). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Clarendon Press.

Icard, T. (2016). Subjective Probability as Sampling Propensity. In *Review of Philosophy and Psychology* (Vol. 7, Issue 4).

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Ingold, T. (2000). *The perception of the environment: essays on livelihood, dwelling and skill*. Routledge.

Ingold, T. (2001). From the transmission of representations to the education of attention. In H. Whitehouse (Ed.), *The debated mind: Evolutionary psychology versus ethnography* (pp. 113–153). Berg.

Jacobs, D. M., & Michaels, C. F. (2007). Direct Learning. *Ecological Psychology*, *19*(4), 321–349.

Jarecki, J. B., Tan, J. H., & Jenny, M. A. (2020). A framework for building cognitive process models. *Psychonomic Bulletin & Review*, *27,* 1218–1229.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026.

Johnson-Laird, P. N., & Khemlani, S. S. (2017). *Mental Models and Causation* (M. R. Waldmann (Ed.); Vol. 1, pp. 169–187). Oxford University Press.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2*. John Wiley & Sons.

Johnson, S. G. B. B., & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, *143*(6), 2223–2241.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A Brief Survey of Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association*, *91*(433), 401–407.

Jurgens, A., & Kirchhoff, M. D. (2019). Enactive Social Cognition: Diachronic Constitution & Coupled Anticipation. *Consciousness and Cognition*, *70*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus, and Giroux.

Kahneman, D, & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahnemand (Eds.), *Heurstics of Intuitive Judgment: Extensions and Applications*. Cambridge University Press.

Katsimpokis, D., Hawkins, G. E., & van Maanen, L. (2020). Not all Speed-Accuracy Trade-Off Manipulations Have the Same Psychological Effect. *Computational Brain & Behavior*, *3*(3), 252–268.

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *64*(1–2), 35–73.

Keuken, M. C., Van Maanen, L., Bogacz, R., Schäfer, A., Neumann, J., Turner, R., & Forstmann, B. U. (2015). The Subthalamic Nucleus During Decision-Making With Multiple Alternatives. *Human Brain Mapping*, *36*(10), 4041–4052.

Khemlani, S. S., & Oppenheimer, D. M. (2011). When One Model Casts Doubt on Another: A Levels-of-Analysis Approach to Causal Discounting. *Psychological Bulletin*, *137*(2), 195–210.

Kiefer, A., & Hohwy, J. (2019). Representation in the Prediction Error Minimization Framework. In J. Symons, P. Calvo, & S. Robins (Eds.), *Routledge Handbook to the Philosophy of Psychology*.

Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, *195*(6), 2519–2540.

Kirfel, L., Icard, T., & Gerstenberg, T. (2021). Inference From Explanation. *Journal of Experimental Psychology: General*.

Kirfel, L., & Lagnado, D. (2018). Statistical norm effects in causal cognition. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, *August*, 615–620.

Kiverstein, J., & Rietveld, E. (2018). Reconceiving representation-hungry cognition: an ecological-enactive proposal. *Adaptive Behavior*, *26*(4), 147–163.

Kiverstein, J., & Rietveld, E. (2021). Scaling-Up Skilled Intentionality to Linguistic Thought. *Synthese*, *198*, 175–194.

Kiverstein, J., van Dijk, L., & Rietveld, E. (2019). The field and landscape of affordances: Koffka's two environments revisited. *Synthese*, *198*, 2279–2296.

Klaassen, P., Rietveld, E., & Topal, J. (2010). Inviting complementary perspectives on situated normativity in everyday life. *Phenomenology and the Cognitive Sciences*, *9*(1), 53–73.

Kleinjans, K. J., & van Soest, A. (2014). Rounding, focal point answers, and nonresponse to subjective probability questions. *Journal of Applied Econometrics*, *29*, 567–585.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Kocher, M. G., & Sutter, M. (2006). Time is money-Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization*, *61*, 375–392.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Kolvoort, I. R., Davis, Z. J., van Maanen, L., & Rehder, B. (2021). Variability in Causal Reasoning. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

Kolvoort, I. R., Fisher, E., Van Rooij, R., Schulz, K., & Van Maanen, L. (2023). Probabilistic causal reasoning under time pressure. *Under Review*.

Kolvoort, I. R., & Rietveld, E. (2022). Affordances for Situating the Embodied Mind in Sociocultural Practice. In Z. Djebbara (Ed.), *Affordances in Everyday Life*.

Kolvoort, I. R., Schulz, K., & Rietveld, E. (2023). The Causal Mind: An Affordance-Based Account of Causal Engagement. *Adaptive Behavior*.

Kolvoort, I. R., Temme, N., & Van Maanen, L. (2023). The Bayesian Mutation Sampler explains distributions of causal judgments. *Open Mind*.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior. *Educational and Psychological Measurement*, *67*(4), 606–619.

Kruis, J., Maris, G., Marsman, M., Bolsinova, M., & van der Maas, H. L. J. (2020). Deviations of rational choice: an integrative explanation of the endowment and several context effects. *Scientific Reports*, *10*, 16226.

Krynski, T. R., & Tenenbaum, J. B. (2007). The Role of Causality in Judgment Under Uncertainty. *Journal of Experimental Psychology: General*, *136*(3), 430–450.

Kuhn, D. (1989). Children and Adults as Intuitive Scientists. *Psychological Review*, *96*(4), 674–689.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*, *40*.

Lee, H. S., & Holyoak, K. J. (2008). The Role of Causal Models in Analogical Inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1111–1122.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust Modeling in Cognitive Science. *Computational Brain and Behavior*, *2*(3–4), 141–153.

Levin, D. A., & Peres, Y. (2017). *Markov Chains and Mixing Times*. American Mathematical Society.

Lewis, D. (1974). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, *25*, 2690–2798.

Lin, Y.-S., Heathcote, A., & Holmes, W. R. (2019). Parallel probability density approximation. *Behavior Research Methods*, *51*, 2777–2799.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., Holyoak, K. J., & Org, E. (2008). Bayesian Generic Priors for Causal Learning. *Psychological Review*, *115*(4), 955–984.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502.

Maaß, S. C., De Jong, J., Van Maanen, L., & Van Rijn, H. (2021). Conceptually plausible Bayesian inference in interval timing. *Royal Society Open Science*, *8*.

Mann, H. B., & Whitney, D. R. (1957). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.

Marchant, M., Toro-Hernandez, F., & Chaigneau, S. E. (2021). Know your priors: Task specific priors reflect subjective expectations in Bayesian models of categorization. *PsyArXiv*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*.

Maule, A. J., Robert, G., Hockey, J., Bdzola, L., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: Changes in affective state and information processing strategy. *Acta Psychologica*, *104*(3), 283–301.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and Causes: Dispositional Intuitions As a Guide to Causal Structure. *Cognitive Science*, *39*(1), 65–95.

McGann, M., De Jaegher, H., & Di Paolo, E. (2013). Enaction and Psychology. *Review of General Psychology*, *17*(2), 203–209.

McGill, A. L. (1995). American and Thai Managers′ Explanations for Poor Company Performance: Role of Perspective and Culture in Causal Selection. *Organizational Behavior and Human Decision Processes*, *61*, 16–27.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, *15*(1), 75–80.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3).

Meder, B., & Mayrhofer, R. (2017). Diagnostic causal reasoning with verbal information. *Cognitive Psychology*, *96*, 54–84.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277–301.

Menzies, P., & Price, H. (1993). Causation as a Secondary Quality. *The British Journal for the Philosophy of Science*, *44*(2), 187–203.

Merleau-Ponty, M. (1962). *Phenomenology of Perception*. Routledge.

Merleau-Ponty, M. (2003). *Nature: Course Notes from the College de France*. Northwestern University Press.

Mestdagh, M., Verdonck, S., Meers, K., Loossens, T., & Tuerlinckx, F. (2019). Prepaid parameter estimation without likelihoods. In *PLoS Computational Biology* (Vol. 15, Issue 9). Public Library of Science.

Michotte, A. E. (1963). *The perception of causality*. Methuen & Co.

Miletić, S., & van Maanen, L. (2019). Caution in decision-making under time pressure is mediated by timing ability. *Cognitive Psychology*, *110*, 1–35.

Mistry, P. K., Pothos, E. M., Vandekerckhove, J., & Trueblood, J. S. (2018). A quantum probability account of individual differences in causal reasoning. *Journal of Mathematical Psychology*, *87*, 76–97.

Morey, R. D., & Rouder, J. N. (2014). *BayesFactor package for R* (0.9.12-4.3).

Morris, M. W., & Larrick, R. (1995). When One Cause Casts Doubt On Another: A Normative Analysis Of Discounting In Causal Attribution. *Psychological Review*, *102*(2).

Morris, M. W., Nisbett, R., & Peng, K. (1995). Causal attribution across domains and cultures. In D. Sperber, D. Premack, & A. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 577–613). Oxford University Press.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin and Review*, *15*(3), 465–494.

Muentener, P., & Bonawitz, E. (2017a). The development of causal reasoning. In *The Oxford Handbook of Causal Reasoning*.

Muentener, P., & Bonawitz, E. (2017b). The Development of Causal Reasoning. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 2–41). Oxford University Press.

Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences - a model-based review. *Neuroscience*, *277*, 872–884.

Müller, H., & Sternad, D. (2004). Decomposition of Variability in the Execution of Goal-Oriented Tasks: Three Components of Skill Improvement. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(1), 212–233.

Myin, E. (2016). Perception as something we do. *Journal of Consciousness Studies*, *23*(5–6), 80–104.

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203.

Noë, A. (2004). *Action in perception*. MIT press.

Noë, A. (2012). *Varieties of Presence*. Harvard University Press.

O'Neill, K., Henne, P., Bello, P., Pearson, J., & De Brigard, F. (2022). Confidence and gradation in causal judgment. *Cognition*, *223*, 105036.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 305–330.

Oppenheimer, D. M. (2004). Spontaneous Discounting of Availability in Frequency Judgment Tasks. *Psychological Science*, *15*(2), 100–105.

Oppenheimer, D. M., Tenenbaum, J. B., & Krynski, T. R. (2013). Categorization as Causal Explanation. Discounting and Augmenting in a Bayesian Framework. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 58, pp. 203–231). Academic Press Inc.

Ordóñez, L., & Benson, L. (1997). Decisions under Time Pressure: How Time Constraint Affects Risky Decision Making. *Organizational Behavior and Human Decision Processes*, *71*(2), 121–140.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433.

Paré, D., & Cree, G. (2009). Web-based image norming: How do object familiarity and visual complexity ratings compare when collected in-lab versus online? *Behavior Research Methods*, *41*(3), 699–704.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186–216.

Park, J., & Sloman, S. A. (2014). Causal explanation in the face of contradiction. *Memory and Cognition*, *42*(5), 806–820.

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, *102*(November 2017), 127–144.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Peterson, C. R., & Beach, L. R. (1967). Man As an Intuitive Statistician. *Psychological Bulletin*, *68*(1), 29–46.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354.

Phillips, L. D., Hays, W. L., & Edwards, W. (1966). Conservatism in Complex Probabilistic Inference. *IEEE Transactions on Human Factors in Electronics*, *HFE-7*(1), 7–18.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901.

Price, H. (2017). Causation, Intervention, and Agency. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation* (pp. 73–98). Oxford University Press.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Kothe, U. (2022). BayesFlow: Learning Complex Stochastic Models With Invertible Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(4), 1452–1466.

Radev, S. T., Wieschen, E. M., & Voss, A. (2020). Amortized Bayesian Inference for Models of Cognition. *ArXiv Preprint*.

Rahnev, D., Desender, K., Lee, A. L. F. F., Adler, W. T., Aguilar-Lleyda, D., Soto, D., Sun, S., Van Boxtel, J. J. A. A., Wang, S., Weidemann, C. T., Weindel, G., Wierzchoń, M., Xu, X., Ye, Q., Yeon, J., Zou, F., Zylberberg, A., Akdoğan, B., Arbuzova, P., … Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719.

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, *13*(4), 626–635.

Ratcliff, R., Zandt, T. Van, & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General*, *124*(4), 352–374.

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

Rehder, B. (2018). Beyond Markov: Accounting for independence violations in causal reasoning. *Cognitive Psychology*, *103*(January), 42–84.

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

Rehder, B., & Davis, Z. J. (2021). Testing a Process Model of Causal Reasoning With Inhibitory Causal Links. *Proceedings of the Annual Meeting of the Cognitive Science Society 43*, *43*.

Rehder, B., Davis, Z. J., & Bramley, N. (2022). The Paradox of Time in Dynamic Causal Systems. *Entropy*, *24*(7).

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

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, *101*(2), 900–926.

Ridderinkhof, K. R. (2002). Micro- and macro-adjustments of task set: Activation and suppression in conflict tasks. *Psychological Research*, *66*(4), 312–323.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer & P. M. Todd (Eds.), *Evolution and cognition. Simple heuristics that make us smart* (pp. 141–167). Oxford University Press.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*(2), 258–276.

Rietveld, E. (2008). Situated normativity: The normative aspect of embodied cognition in unreflective action. *Mind*, *117*(468), 973–997.

Rietveld, E. (2012). Bodily intentionality and social affordances incontext. In F. Paglieri (Ed.), *Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness* (pp. 207–226). J. Benjamins.

Rietveld, E. (2022). Change-Ability for a World in Flux. *Adaptive Behavior*.

Rietveld, E., & Brouwers, A. A. (2017). Optimal grip on affordances in architectural design practices: an ethnography. *Phenom Cogn Sci*, *16*, 545–564.

Rietveld, E., Denys, D., & Van Westen, M. (2018). Ecological-Enactive Cognition as Engaging with a Field of Relevant Affordances: The Skilled Intentionality Framework (SIF). In *The Oxford Handbook of 4E Cognition* (pp. 41–70).

Rietveld, E., & Kiverstein, J. (2014). A Rich Landscape of Affordances. *Ecological Psychology*, *26*(4), 325–352.

Rottman, B. M. (2017). The Acquisition and Use of Causal Structure Knowledge. In M. R. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning*. Oxford University Press.

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

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

Rubinstein, A. (2007). Instinctive and Cognitive Reasoning: A study of Response Times. *The Economic Journal*, *117*, 1243–1259.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176.

Sanborn, A. N., & Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Sanborn, A. N., Heller, K., Austerweil, J. L., & Chater, N. (2021). REFRESH: A new approach to modeling dimensional biases in perceptual similarity and categorization. *Psychological Review*, *128*(6), 1145–1186.

Scarantino, A. (2003). Affordances Explained. *Philosophy of Science*, *70*(5), 949–961.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times With a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, *34*(3), 213–232.

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and Accuracy. In *Acta Psychologica* (Vol. 27). North-Holland Publishing Co.

Schwarz, G. (1978). Estimating the Dimensions of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*(2), 175–192.

Shapiro, L., & Spaulding, S. (2021). Embodied cognition. In E N Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2021 Edition)*.

Sheather, S. J., & Jones, M. C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683–690.

Sloman, S. A. (2005). *Causal models: how people think about the world and its alternatives*. Oxford University Press.

Sloman, S. A. (2009). *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press.

Sloman, S. A., & Lagnado, D. (2015). Causality in Thought. *Annual Review of Psychology*, *66*, 223–247.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*(2–3), 167–196.

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*(1), B1–B11.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.

Stanovich, K. E. (1999). Who Is Rational? In *Who Is Rational?* Psychology Press.

Stepp, N., & Turvey, M. T. (2015). The Muddle of Anticipation. *Ecological Psychology*, *27*(2), 103–126.

Stewart, J. (2010). Foundational issues in enaction as a paradigm for cognitive science: From the origin of life to consciousness and writing. In J. Stewart, O. Gapenna, & E. A. Di Paolo (Eds.), *Enaction: Toward a new paradigm fo cognitive science*. MIT press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Stoffregen, T. A. (2003). Affordances as Properties of the Animal-Environment System. *Ecological Psychology*, *15*(2), 115–134.

Sussman, A. B., & Oppenheimer, D. (2011). A Causal Model Theory of Judgment. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, *33*.

Szul, M. J., Bompas, A., Sumner, P., & Zhang, J. (2020). The validity and consistency of continuous joystick response in perceptual decision-making. *Behavior Research Methods*, *52*(2), 681–693.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410–441.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285.

Tipper, S. P., Paul, M. A., & Hayes, A. E. (2006). Vision-for-action: The effect of object property discrimination and action state on affordance compatibility states. *Psychonomic Bulletin & Review*, *13*(3), 493–498.

Trueblood, J. S., Yearsley, J. M., & Pothos, E. M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experimental Psychology: General*, *146*(9), 1307–1341.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 65–79.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin and Review*, *21*(2), 227–250.

Turvey, M. T. (1992). Affordances and Prospective Control: An Outline of the Ontology. *Ecological Psychology*, *4*(3), 173–187.

Turvey, M. T., Shaw, R. E., Reed, E. S., & Mace, W. M. (1981). Ecological laws of perceiving and acting: In reply to Fodor and Pylyshyn (1981). *Cognition*, *9*, 237–304.

Tversky, A., & Kahneman, D. (1982). Causal Schemata in judgments under uncertainty. In Daniel Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 117–128). Cambridge University Press.

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Van Den Herik, J. C. (2018). Attentional actions- A n ecological-enactive account of utterances of concrete words. *Psychology of Language and Communication*, *22*(1), 90–123.

Van Den Herik, J. C., & Rietveld, E. (2021). Reflective Situated Normativity. *Philosophical Studies*, *178*(10), 3371–3389.

van Dijk, L., & Kiverstein, J. (2021). Direct perception in context: radical empiricist reflections on the medium. *Synthese*, *198*, 8389–8411.

van Dijk, L., & Rietveld, E. (2017). Foregrounding sociomaterial practice in our understanding of affordances: the Skilled Intentionality Framework. *Frontiers in Psychology*, *7*.

van Dijk, L., & Rietveld, E. (2020). Situated imagination. *Phenomenology and the Cognitive Sciences*, 1–23.

van Dijk, L., & Rietveld, E. (2021a). Situated anticipation. *Synthese*, *198*, 349–371.

van Dijk, L., & Rietveld, E. (2021b). Situated talking. *Language Sciences*, *87*, 101389.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E. J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics*, *47*(16), 2984–3006.

van Maanen, L. (2016). Is There Evidence for a Mixture of Processes in Speed-Accuracy Trade-Off Behavior? *Topics in Cognitive Science*, *8*(1), 279–290.

Van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural Correlates of Trial-to-Trial Fluctuations in Response Caution. *Journal of*

Neuroscience, *31*(48), 17488–17495.

Van Maanen, L., Couto, J., & Lebreton, M. (2016). Three boundary conditions for computing the fixed-point property in binary mixture data. *PLoS ONE*, *11*.

Van Maanen, L., De Jong, R., & Van Rijn, H. (2014). How to assess the existence of competing strategies in cognitive tasks: A primer on the fixed-point property. *PLoS ONE*, *9*(8).

van Maanen, L., Grasman, R. P. P. P. P. P., Forstmann, B. U., Keuken, M. C., Brown, S. D., & Wagenmakers, E.-J. J. (2012). Similarity and number of alternatives in the random-dot motion paradigm. *Attention, Perception, and Psychophysics*, *74*(4), 739–753.

van Maanen, L., Katsimpokis, D., & van Campen, A. D. (2018). Fast and slow errors: Logistic regression to identify patterns in accuracy–response time relationships. *Behavior Research Methods*, *51*(5), 2378–2389.

Van Maanen, L., & Miletić, S. (2021). The interpretation of behavior-model correlations in unidentified cognitive models. *Psychonomic Bulletin & Review*, *28*, 374–383.

van Maanen, L., van der Heiden, R. M., Bootsma, S. T., & Janssen, C. P. (2021). Identifiability and Specificity of the Two-Point Visual Control Model of Steering. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Van Maanen, L., & Van Rijn, H. (2010). The Locus of the Gratton Effect in Picture–Word Interference. *Topics in Cognitive Science*, *2*(1), 168–180.

van Ravenzwaaij, D., Brown, S. D., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*(3), 381–393.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin and Review*, *25*(1), 143–154.

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*(6), 463–473.

Van Rijn, H., Someren, M. van, & Maas, H. van der. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257.

van Westen, M., Rietveld, E., & Denys, D. (2019). Effective deep brain stimulation for obsessive-compulsive disorder requires clinical expertise. *Frontiers in Psychology*, *10*(OCT), 2294.

van Westen, M., Rietveld, E., van Hout, A., Denys, D., & van Westen mvanwesten, M. (2021). “Deep brain stimulation is no ON/OFF-switch”: an ethnography of clinical expertise in psychiatric practice. *Phenomenology and the Cognitive Sciences*.

Varela, F. J., Rosch, E., & Thompson, E. (1991). *The embodied mind: Cognitive Science and Human Experience*. MIT press.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable Causal Relationships Are Better Causal Relationships. *Cognitive Science*, *42*(4), 1265–1296.

Verdonck, S., & Tuerlinckx, F. (2013). Factoring out non-decision time in choice RT data: Theory and implications. *Psychological Review*, *123*(2).

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*(2), 179–197.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Waldmann, M. R. (2017a). Causal Reasoning: An Introduction. In *Oxford Handbook of Causal Reasoning*.

Waldmann, M. R. (2017b). *The Oxford Handbook of Causal Reasoning* (Michael R Waldmann (Ed.)).

Waldmann, M R, Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal Learning in Rats and Humans: A Minimal Rational Model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian cognitive science* (pp. 453–484). Oxford University Press.

Waldmann, M R, & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*(1), 27–58.

Waldmann, M R, & Hagmayer, Y. (2013). Causal Reasoning. In *The Oxford Handbook of cognitive Psychology* (Issue November, pp. 1–24). Oxford University Press.

Waldmann, M R, & Mayrhofer, R. (2016). Hybrid Causal Representations. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 65, pp. 85–127). Elsevier Ltd.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. *Management Science*, *39*(2), 176–190.

Walsh, C. R., & Sloman, S. A. (2011). The Meaning of Cause and Prevent: The Role of Causal Mechanism. *Mind and Language*, *26*(1), 21–52.

Walsh, C. R., & Sloman, S. A. (2005). The Meaning of Cause and Prevent: The Role of Causal Mechanism. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *27*, 27.

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 683–703.

Welsh, M. B., & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, *119*(1), 1–14.

White, C. N., Servant, M., & Logan, G. D. (2015). *Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study*.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and informaiton processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, 1–33.

Withagen, R., de Poel, H. J., Araújo, D., & Pepping, G. J. (2012). Affordances can invite behavior: Reconsidering the relationship between affordances and agency. *New Ideas in Psychology*, *30*(2), 250–258.

Wittgenstein, L. (1953). *Philosophical investigations* (G.E.M. Ans). Blackwell Publishing Ltd.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2011). A Philosopher Looks at Tool Use and Causal Understanding. In T. McCormack, C. Hoerl, & S. Butterfill (Eds.), *Tool use and causal cognition* (pp. 18–50). Oxford University Press.

Woodward, J. (2014). A functional account of causation; or, a defense of the legitimacy of causal thinking by reference to the only standard that matters—usefulness (as opposed to metaphysics or agreement with intuitive judgment). *Philosophy of Science*, *81*(5), 691–713.

Woodward, J. (2016). Causation and Manipulability. In Edward N Zalta (Ed.), *Stanford Encyclopedia of Philosophy (Winter 2016 Edition)* (pp. 1–31). Metaphysics Research Lab, Stanford University.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1310–1321.

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29.

Yitzhaki, S. (2003). Gini's Mean Difference: A superior Measure of Variability for Non-Normal Distributions. *METRON*, *61*(2), 285–316.

Young, D. L., Goodie, A. S., Hall, D. B., & Wu, E. (2012). Decision making under time pressure, modeled in a prospect theory framework. *Organizational Behavior and Human Decision Processes*, *118*(2), 179–188.

Zahidi, K., & Myin, E. (2016). Radically enactive numerical cognition. In G. Etzelmuller & C. Tewes (Eds.), *Embodiment in evolution and culture* (pp. 57–71).

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian Sampler: Generic Bayesian Inference Causes Incoherence in Human Probability Judgments. *Psychological Review*, *127*(5), 719.