

Answering Causal Queries About Singular Cases

An Evaluation of a New Computational Model

Dissertation

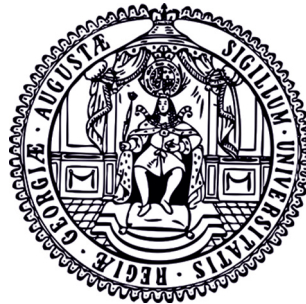
Zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

im Promotionsprogramm Behavior and Cognition (BeCog)

der Georg-August University School of Science (GAUSS)



vorgelegt von

Simon Stephan

aus Hann. Münden (Hedemünden)

Göttingen, 2019

Betreuungsausschuss

Prof. Dr. Michael R. Waldmann, Kognitionswissenschaft und
Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität
Göttingen

Prof. Dr. Hannes Rakoczy, Kognitive Entwicklungspsychologie,
Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. Markus Steinbach, Seminar für Deutsche Philologie, Universität
Göttingen

Mitglieder der Prüfungskommission

Referent: Prof. Dr. Michael R. Waldmann, Kognitionswissenschaft und
Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität
Göttingen

Korreferent: Prof. Dr. Hannes Rakoczy, Kognitive Entwicklungspsychologie,
Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Dr. Tanya Behne, Kognitive Entwicklungspsychologie, Georg-Elias-Müller-Institut
für Psychologie, Universität Göttingen

Prof. Dr. Nivedita Mani, Forschungsgruppe Sprachpsychologie,
Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. Annekathrin Schacht, Affektive Neurowissenschaft und
Psychophysiologie, Georg-Elias-Müller-Institut für Psychologie, Universität
Göttingen

Prof. Dr. Markus Steinbach, Seminar für Deutsche Philologie, Universität
Göttingen

Tag der mündlichen Prüfung: 28.02.2019

To my grandfather
(1933 – 2018)
and my parents

Acknowledgements

This thesis marks the end of an intense and exciting period of my life. I learned a lot during this time. Above all else, I learned how fascinating cognitive science is and how incredibly lucky I must consider myself for having had the opportunity to work with and learn from so many inspiring people. I want to express my deepest gratitude to them.

The first person I want to thank is my mentor and supervisor Michael Waldmann. With unparalleled inner calmness and an extremely keen mind, he has supported and encouraged me throughout the whole time of my PhD in the most constructive manner I can think of. He has given me a great deal of freedom to think about and explore the things that really interested me. I am deeply grateful to him and I am looking forward to continuing our collaboration.

Thanks to Hannes Rakoczy and Markus Steinbach for having been in my thesis committee, and to Annekathrin Schacht, Tanya Behne, and Nivedita Mani for agreeing to be members of the examination board.

The person who inspired me to start a PhD was Jonas Nagel, with whom I started working together when I was still in my bachelor's program. Jonas introduced me to the world of cognitive science and ignited my enthusiasm for it. His ability to think through even the deepest philosophical problems and to creatively turn them into experimental ideas has always fascinated me. Thanks, Jonas, for all your advice and for the great time we spent together.

I want to thank my colleagues York Hagmayer and Alex Wiegmann, who were always interested a lot in what I was doing and who always gave me valuable feedback and advice. A special thanks goes to Ralf Mayrhofer, whose office door was always open for me. I benefited a lot from his ingenious mind. Learning from and working with him was a great pleasure. Thanks also to my other colleagues and friends Neele Engemann, Juan Marulanda, and Christopher Pohr for your mental support during the last phase of my PhD. Thanks also to Pascale Willemsen and to Tobias Gerstenberg, with whom I had the privilege to work together and to discuss many interesting topics about causality.

I would like to express my gratitude to the Cognitive Science Society for having awarded me the Computational Modeling Prize. I also want to thank the DFG for having provided funding for my position and the experiments.

Thank you, Sarah, for all your support, your smartness, your warmth, your curiosity, your ideas, and your love. Life is shining bright with you.

Finally, I would like to thank my parents, Siegmund and Verena, and my sister, Juliane. You have been my safe haven throughout my life. Thanks for all your love and support.

Preliminary Note

The present thesis is a publication-based (cumulative) dissertation. It is based on the following two original research articles that have been published in or are currently submitted to international peer-reviewed journals.

Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, 10, 242 – 257.

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (submitted). The role of causal strength and temporal information in singular causation judgments. *Manuscript submitted for publication*.

In the present thesis, I will present the new model that I developed and empirically tested across these two articles, together with the relevant theoretical background. I will then summarize the main empirical findings from the two articles and provide an extended general discussion. All parts of the dissertation were written by me and assistance of third parties was only accepted if it was scientifically justifiable and acceptable in regards to the examination regulations. All sources have been quoted.

The original articles are reprinted in Appendices B and C. I served as first author in both articles. In particular, I was responsible for (a) developing the theory, (b) designing and conducting the experiments, (c) analyzing and interpreting the data, and (d) writing up and publishing the manuscripts. Both articles have precursors in the form of peer-reviewed proceedings papers that I wrote during my PhD and which have been published by the Cognitive Science Society. These proceedings papers can be found in Appendices D, E, and F. Apart from this main project I also worked on additional projects during my PhD, about predictive causal reasoning in interpolated causal chains, the role of mechanisms in causal explanations, and causal judgments in the context of omissions. The articles that resulted from these projects are included in Appendices G to J, and brief summaries of these additional articles are given in Appendix A.

Abstract

This thesis addresses the question of how causal queries about singular cases can be answered. Singular causation queries refer to causal connections between actually occurred events that can be localized in space and time. “Did the storm last night cause the flower pot to break apart?” or “Was it the gunshot instead of the poisoning that caused the victim’s death?” are vivid examples. Singular causation queries can be contrasted with general causation queries, which refer to causal connections between re-instantiable types (e.g., “Does smoking cause lung cancer?”). Singular causation queries are prevalent in our daily lives and important in many professional disciplines, such as the law, medicine, or engineering. But how can we assess whether an event c caused another event e ? Given that causal connections are not directly perceivable, how can a causal co-occurrence of events be discriminated from a mere *coincidental* one? In this thesis, I propose a new computational model that is intended to help answering this question. The model is based on Cheng and Novick’s (2005) Power Model of Causal Attribution, according to which general-level causal knowledge about the potential causes’ powers (Cheng, 1997) is crucial. The causal power of a cause is the probability with which it brings about the effect. I will argue that more than that is needed: The assessment of singular causation also needs to take *temporal information* into account. By focusing solely on causal power, Cheng and Novick’s model assigns singular causal responsibility to a target cause whenever the cause was sufficiently powerful for the effect. It thus fails to take into consideration that causes can be *preempted* in their efficacy by alternative causes. To account for the problem of causal preemption, the new model combines information about *causal power* and *temporal information*. Two different types of temporal information are identified as relevant: information about *causal latency*, which is the time it takes a cause to unfold its power, and information about the *onset difference* between the potential causes. A second problem of the original model is that it relies solely on point estimates of the power parameters, and thus does not take into account that reasoners incorporate uncertainty about the underlying general causal structure and the parameters. To account for this problem of inferential uncertainty, the new model is embedded into the structure induction framework (Meder, Mayrhofer, & Waldmann, 2014). The new model was experimentally tested across two research articles (Stephan, Mayrhofer, & Waldmann, submitted; Stephan & Waldmann, 2018). The experiments revealed that the new model better accounts for people’s singular causation judgments than the original model.

Contents

1	Introduction	1
2	Cheng's (1997) Causal Power Theory	5
2.1	Estimating unobservable causal powers from observable covariations	5
2.2	Evidence for the psychological validity of the theory	11
2.2.1	Zero-contingencies under ceiling vs. non-ceiling conditions	12
2.2.2	Different intuitions towards constant positive probabilistic contrasts	13
2.2.3	Non-interactive causal powers as a tacit default assumption	14
2.3	Problematic findings	17
2.4	Summary	18
3	Applying Causal Power Knowledge to Evaluate Singular Cases – The Power Framework of Causal Attribution	20
3.1	Different equations for different states of knowledge about the target cause	21
3.2	A filter of probabilistic sufficiency	23
3.3	What the framework seems to get right	24
3.3.1	The stronger the target cause the better	24
3.3.2	The less likely and the weaker alternative causes the better – Holmesian inference	24
3.4	The blind spot of the framework – the possibility of causal preemption	25
3.4.1	Preemption and the assessment of causal power	29
3.5	Summary	30
4	A Generalization of the Framework Sensitive to Causal Preemption	32
4.1	A generalized model sensitive to preemption	32
4.2	Temporal information determines the Alpha parameter	33
4.2.1	Onset difference	34
4.2.2	Causal latency	34
4.2.3	Modeling causal latencies	37

4.2.4	A formalization of the Alpha parameter	38
4.2.5	Approximating alpha with a Monte Carlo algorithm	39
4.3	What if alternative causes are unobserved	40
4.4	What if temporal information is not available	42
4.5	Summary	42
5	Incorporating General Causal Structure and Parameter Uncertainty	44
5.1	Incorporating causal parameter uncertainty	46
5.2	Incorporating general-causal structure uncertainty	46
5.3	The Structure Induction Model of Singular Causation Judgments	48
5.3.1	Estimating the general causal structures' probabilities	48
5.3.2	Estimating each potential structure's set of parameters	49
5.3.3	Estimating the probability of singular causation under each parametrized structure	51
5.3.4	Obtaining a single estimate for singular causation	52
5.4	Summary	53
6	Summary of the Empirical Findings	55
6.1	Experiments of Stephan and Waldmann (2018)	56
6.1.1	Experiments 1a and 1b	56
6.1.2	Experiment 2	59
6.1.3	Summary	63
6.2	Experiments of Stephan, Mayrhofer, and Waldmann (submitted)	63
6.2.1	Overview of cover story, procedure, and materials	64
6.2.2	Experiment 1	66
6.2.3	Experiment 2	68
6.2.4	Experiment 3	70
6.2.5	Experiment 4	73
6.2.6	Summary	75
7	General Discussion	77
7.1	Empirical limitations	78
7.2	General theoretical considerations and outlook	80
7.2.1	The role of knowledge about causal mechanisms	80
7.2.2	Assessing singular causation vs. selecting singular causes	82
7.2.3	Singular events and their corresponding types	84

7.2.4	Assessing the relevant parameters cross-sectionally vs. longi- tutinally	85
7.3	Conclusion	86
A	Summaries of the Additional Articles	i
A.1	Summary of Stephan, Tentori, Pighin, and Waldmann (in prep) . . .	i
A.2	Summary of Nagel and Stephan (2015, 2016)	ii
A.3	Summary of Stephan, Willemsen, and Gerstenberg (2017)	iii
B	Stephan and Waldmann (2018)	v
C	Stephan, Mayrhofer, and Waldmann (subm.)	xxii
D	Stephan and Waldmann (2016)	lxxviii
E	Stephan and Waldmann (2017)	lxxxv
F	Stephan, Mayrhofer, and Waldmann (2018)	xcii
G	Stephan, Tentori, Pighin, and Waldmann (in preparation)	xcix
H	Nagel and Stephan (2015)	clxii
I	Nagel and Stephan (2016)	clxix
J	Stephan, Willemsen, and Gerstenberg (2017)	clxxvi
	Curriculum Vitae	clxxxiii

List of Figures

2.1	Results of a fictitious medical study	7
2.2	The common-effect causal structure assumed by the Causal Power Theory	8
2.3	Illustration of how the Causal Power Theory explains the observed contingency	9
2.4	Euler diagram illustrating how the causal power theory explains the observable ΔP	11
2.5	Different situations in which the contingency is zero	13
2.6	Different situations in which the contingency is constant at a positive value	14
2.7	The two different contingency data sets that Liljeholm and Cheng (2007) presented	15
3.1	Illustration of the principle behind Equation 3.2	23
3.2	Neuron diagram modeling a situation of causal preemption	26
3.3	The singular causal structures compatible with the general causal structure and with a singular observation	28
3.4	Illustration showing that the preemptive relation between target cause and alternative causes is conceptually unproblematic for the estimation of causal power.	29
4.1	Illustration of the temporal components relevant for the assessment of the preemptive relation	35
4.2	Example of gamma distributions	38
4.3	Example of gamma and exponential distributions	41
5.1	The two causal structures assumed by the Structure Induction Framework	47
5.2	The two core inferential steps of the SI Model	50
6.1	Results of Experiment 1a (Stephan & Waldmann, 2018)	58
6.2	Predictions and results of Experiment 2 (Stephan & Waldmann, 2018)	61

6.3	Illustration of the learning task used in Experiment 1 in Stephan et al. (submitted).	64
6.4	Pairs of gamma distributions contrasted in the five conditions of Exp. 2 in Stephan et al. (submitted)	68
6.5	Model predictions and results of Experiment 2 in Stephan et al. (submitted)	70
6.6	Model predictions and results of Experiment 3 in Stephan et al. (submitted)	72
6.7	Model predictions and results of Experiment 4 in Stephan et al. (submitted)	75
7.1	Illustration showing the difference between the concepts of singular causation and causal selection.	83

List of Tables

2.1	The contingency table	6
5.1	Different sets of observations for which the Power Framework of Causal Attribution predicts maximum values of $P(c \rightarrow e c, e)$ because the effect is sometimes observed in the target cause's presence but never in its absence.	45

Chapter 1

Introduction

“Shallow men [and women] believe in luck or in circumstance. Strong men [and women] believe in cause and effect.”

- Ralph Waldo Emerson

Our mind possesses the remarkable capacity to structure the world into causes and effects. We do not have the feeling that we live in a world consisting of pure chaos, in which things *just happen*, where events pop up out of nothingness and bootstrap themselves into existence. Rather, we think that things happen because there are causes at play that make them happen. Lung disease rates increased dramatically in the first half of the twentieth century because more and more people had fallen prey to the seductive tobacco commercials and started to light up. The high density of shooting stars whooshing over the night sky every August is a celestial spectacle that we can observe because our earth revolves around the sun on an orbit that intersects the asteroid belt of our solar system. The orbits, in turn, on which all planets of the solar system are traveling exist because the sun with its gigantic mass is curving space and time around it. Moderate exercising helps improve (or maintain) good health. Drinking a bottle of red wine causes headaches. The space ship Columbia and its seven crew members were lost because a tiny piece of insulating foam that had broken loose from the left bipod ramp section during launch teared a small hole in the thermal protection shield of the orbiter's left wing.

The ability to reason causally about how the world works is regarded by cognitive scientists as one of the core elements of our thinking and reasoning (see Sloman, 2005; Waldmann, 2017; Waldmann & Hagmayer, 2013). Thinking causally allows us to predict future events, to make diagnoses, to intervene effectively, and to explain. It enables us to exert control over our environment and allows us to adapt to it more successfully than probably any other species.

The question of this thesis is how causal judgments about *singular cases* can be made. By singular cases singular co-occurrences of events are meant that can be located in space and time, like the breaking loose of a tiny piece of insulating foam that occurred on January 16, 2003, and the destruction of the Columbia that occurred two weeks later. *Singular causation* queries can be contrasted with *general* or generic causal queries, which refer to causal relationships holding between re-instantiable *types of events* or *factors*. Queries like “Does eating red meat cause colon cancer?”, “Do physical exercises increase psychological well-being?” or “Can a particular new medicine cause headaches as a side effect?” are all questions located on the generic level, whereas queries like “Did Peter’s eating red meat cause his colon cancer?”, “Was it my exercising regularly that led to my improved well-being”, or “Was this dose of medication I took responsible for my headaches?” all are concerned with causality in singular cases.

People ask and answer singular causation queries all the time. As children we ask why the dinosaurs got extinct or why one of our teeth began to wobble. As adults we ask why we got this letter from the tax office, or whether we would have maintained a healthy cholesterol level if we had stuck to a low-fat diet. Pinpointing singular causes is important because it helps identifying target points for interventions that allow us to change adverse states into more favorable ones, or to maintain states that we consider to be good for us. Further, providing explanations for events that happened seems to be in its own right an intrinsically gratifying activity for us humans, and singular causes often seem to be the bearers of explanatory relevance (see also Gopnik, 2000; Lombrozo & Vasilyeva, 2017; Strevens, 2008).

Although singular causation queries are ubiquitous in both our mundane lives and in professional disciplines such as the law (see Hart & Honoré, 1959/1985; Lagnado & Gerstenberg, 2017; Russo & Williamson, 2011) or medicine, the question of how it can be established whether two singular events c and e were causally connected or not turns out to be difficult to answer. The reason is, as has been famously postulated by the philosopher David Hume (Hume, 1748/1977), that causal connections between events are not directly observable (though see, e.g., White, 2017, for a different view). How, then, can a *causal* singular co-occurrence of events be distinguished from a mere *coincidental* one?

The central idea is that judgments about singular causation must rely on knowledge about general causation, which, according to dependency theories of causal inference (see Waldmann & Hagmayer, 2013, for an overview), needs to be inferred on the basis of observed patterns of covariation. For example, according to probabilistic theories of causality (e.g., Eells, 1991), we can infer (if the right conditions

are met) that smoking causes lung cancer because the probability of lung cancer is higher among smokers than among non-smokers. However, merely knowing that smoking can *generally* cause lung cancer is clearly not sufficient to conclude that, for instance, Peter's smoking caused his lung cancer. The general relationship between smoking and lung cancer is a rather noisy one. Not all smokers end up with lung cancer and lung cancer can also be caused by other factors (e.g., asbestos). Thus, despite confidence in the existence of a general link between smoking and lung cancer, Peter's singular case could still have been a mere coincidence.

Patricia Cheng (1997) suggested that reasoners would interpret observable patterns of covariation in a particular way that allows to quantify the unobservable strength or *power* of a cause. Causal power represents the probability with which a cause, acting in isolation from alternative causes, brings about the effect. Based on Cheng's seminal Causal Power Theory (1997), Cheng and Laura Novick (2005) have proposed a *Power Framework of Causal Attribution*, according to which knowledge about the powers of causes is the crucial ingredient for the formation of beliefs about singular causation. The power framework of causal attribution provides a set of equations that intend to compute, for different situations in which a reasoner might be, the probability with which a potential cause event c was actually causally connected to a target effect event e , denoted $P(c \rightarrow e)$. In the Bayesian sense, this probability can be understood as the *degree of belief* that c and e were causally connected.

In this thesis, I will argue that the Power Model of Causal Attribution needs to be refined to be successful, and the overall goal of this thesis is to provide this refinement and to present empirical evidence for its psychological validity. One problem of the framework that I will address is of epistemic nature. It has been neglected that singular causation judgments should incorporate uncertainty about the underlying general causal structure and its parameters. Relying on previous work by Griffiths and Tenenbaum (2005) and Meder et al. (2014), I will show how this shortcoming can be remedied through an augmentation of the framework with a Bayesian inference algorithm that quantifies these uncertainties. I will also provide empirical evidence suggesting that lay people's singular causation judgments are indeed sensitive to these inferential uncertainties.

The second problem I will address is in my opinion even more relevant because it is concerned with a *conceptual* rather than a mere epistemic shortcoming of the current framework. A central argument will be that the standard framework has neglected that another type of information is equally crucial for the evaluation of singular causation. This type of information concerns the *temporal relation* between

the potential target cause and potential alternative causes of the effect. I will show that thus far the standard power model of causal attribution identifies a potential cause c that has co-occurred with the effect on an occasion as the singular cause of the observed effect e whenever one can be certain that c had sufficient causal power to produce e . The framework thus neglects the problem that a target cause can be *preempted* in its efficacy by an alternative cause that was also present on the occasion. To deal with the problem of *causal preemption*, I will show that a new term can be added to the equations of the framework, which combines causal power and temporal information to estimate the probability with which the target cause c was preempted on an occasion by an alternative cause a . Several studies that I conducted will demonstrate the validity of the new model.

This thesis is structured as follows: Since a crucial idea is that reasoners must rely on their background knowledge about the powers of the potential causes when they want to assess causality in singular cases, I will in the next chapter introduce Cheng's (1997) famous causal power theory, which explains how the unobservable power of a target cause can be inferred from observable patterns of covariation. Thereafter, in Chapter 3, I will introduce Cheng and Novick's (2005) Power Framework of Causal Attribution, which tells us how knowledge about causal power ought to be applied to answer causal queries about singular cases. I will work out the logic behind the equations of the framework and make clear what I think the framework has already been getting right and, more importantly, what I think it has not been getting right. The focus will then be on how the "what it has not been getting right" can be fixed. I will then propose in Chapters 4 and 5 a refined version of the model that remedies its shortcomings. In Chapter 6, I will summarize the results of the series of experiments that were reported in the two articles on which this dissertation is based, which aimed at testing the psychological validity of the new model. In Chapter 7, I will discuss empirical and theoretical limitations and will point out interesting research questions for future studies.

Chapter 2

Cheng's (1997) Causal Power Theory

The basic idea of the Power Framework of Causal Attribution is that the degree to which a reasoner can believe that an observed singular co-occurrence of events c and e was causal instead of merely coincidental ought to be based on knowledge about the relative causal power with which the corresponding target cause factor C tends to generate the effect in comparison to potential alternative causes A . As causal power is assumed to be an unobservable entity, however, the crucial question that needs to be answered first is how a reasoner can infer the causal power with which a particular cause factor generates its effect. An answer to this question is provided by Patricia Cheng's (1997) Causal Power Theory, which will be introduced in this chapter.

2.1 Estimating unobservable causal powers from observable covariations

The fundamental idea of the Causal Power Theory is that even though the power of a cause is not directly observable, causal powers give rise to and explain observable patterns of covariation (for a philosophical defense of this view see Cartwright, 1989). Importantly, the theory also specifies the conditions under which observable patterns of covariation between a potential target cause factor and an effect factor can be regarded as a valid *diagnostic tool* that can be used to estimate the causal power of the target cause factor.

For an illustration of how the theory works, let us consider a concrete example. Imagine the scenario of a medical study in which the question shall be answered whether a newly developed medicine (C) can cause headaches (E) as a side effect (cf. Buehner, Cheng, & Clifford, 2003; Liljeholm & Cheng, 2007, 2009, who used this scenario in their studies). The standard experimental procedure in this case

Table 2.1: The contingency table

	e	$\neg e$
c	$n(c, e)$	$n(c, \neg e)$
$\neg c$	$n(\neg c, e)$	$n(\neg c, \neg e)$

would be to collect two random samples of patients, one that serves as a treatment group in which the newly developed drug is administered, and another one serving as an untreated control group. The potential cause and the effect factor in this case can be represented as binary variables that can either be instantiated (c or e) or not instantiated ($\neg c$ or $\neg e$). If the cause and effect factor represent binary variables, the causal power of the target cause factor C can be understood as the probability with which it individually tends to generate the effect. The observations of a study involving two binary factors are typically summarized in a 2×2 contingency table as illustrated in Table 2.1. Evidence for the hypothesized generative causal influence of the target cause C is supposed to be signalled in this experimental setting according to normative theories of probabilistic causality (e.g., Eells, 1991; Reichenbach, 1978; Salmon, 1980; Suppes, 1970) if the probability of the effect in the presence of the cause, $P(e|c)$ is greater than the probability of the effect in the absence of the cause, $P(e|\neg c)$. A standard measure that has been proposed in both psychology (Jenkins & Ward, 1965) and philosophy (Salmon, 1965) to quantify the probabilistic dependency between the target cause and the effect is the *contingency* or ΔP model, which is given by the following equation:

$$\Delta P = \frac{n(c, e)}{n(c, e) + n(c, \neg e)} - \frac{n(\neg c, e)}{n(\neg c, e) + n(\neg c, \neg e)} = P(e|c) - P(e|\neg c). \quad (2.1)$$

Imagine our fictitious study had revealed the results that are depicted graphically in Figure 2.1. The left panel shows the treatment group and the right panel shows the untreated control group. As can be seen, out of the twenty-four patients who received the new medicine C twenty ended up with headaches, $n(c, e) = 20$, and four did not develop headaches, $n(c, \neg e) = 4$. The conditional probability of headaches in the presence of C hence is $P(e|c) = \frac{20}{20+4} = \frac{5}{6}$. The control group, by contrast, is split in half. Twelve patients developed headaches, $n(\neg c, e) = 12$ and twelve remained free of headaches, $n(\neg c, \neg e) = 12$, which yields a base rate probability of headaches of $P(e|\neg c) = \frac{12}{12+12} = \frac{1}{2}$. These results thus yield a probabilistic contrast of $\Delta P = P(e|c) - P(e|\neg c) = \frac{20}{24} - \frac{12}{24} = \frac{8}{24} = \frac{1}{3}$. This probabilistic contrast of $\frac{1}{3}$ indicates that the probability of headaches increases by one third in the presence

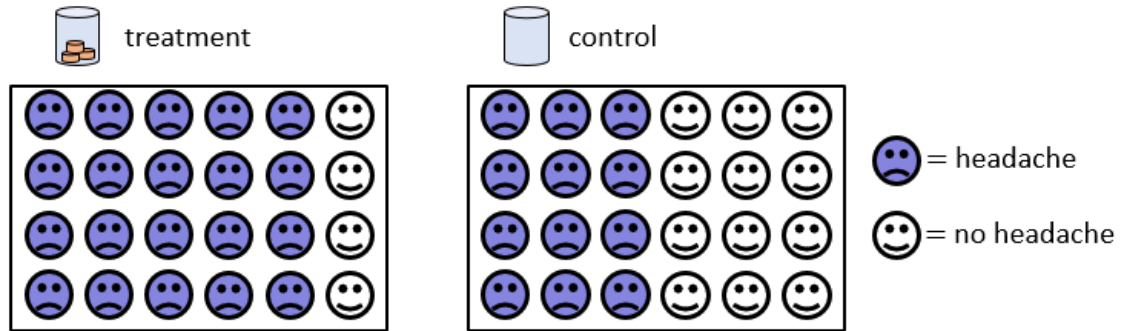


Figure 2.1: Results of a fictitious medical study in which it was investigated whether a newly developed medicine can cause headaches as a side effect. The treatment group is depicted in the left panel. Blue frowning faces indicate the presence of the effect and white smiling faces indicate the absence of the effect.

compared to the absence of the target cause factor.

To see how the Causal Power Theory explains this observed probabilistic contrast by means of causal powers and to see how the causal power of C can be estimated based on these data, we need to consider the assumptions that the theory makes. Even though Cheng (1997) originally has not used any terminology from the causal Bayes nets literature, the assumptions made by her theory have later been shown to instantiate a simple common-effect causal structure in which two cause factors combine their influence independently according to a *noisy logical OR* function (see Glymour, 2003; Griffiths & Tenenbaum, 2005; Pearl, 1988). This causal structure is depicted graphically in Figure 2.2. As can be seen, the effect variable E is assumed to have two potential generative causal sources, C and A . C represents the target cause factor, the medicine in our example, whose causal power shall ultimately be estimated, and A represents the conglomerate of all known and unknown alternative factors that can also generate E (e.g., high blood pressure, dehydration, etc.). C and A are hence assumed to provide an exhaustive set of all potential generative causes of E . In the simplest case, A consists of only one alternative cause. The parameters b_C and b_A attached to the causal arrows that point towards C and A represent the base rates with which C and A occur, respectively. In the given example, the base rate b_C of C is the treatment probability. As half of the subjects were given the pill, this parameter is known to be $b_C = 0.5$. The base rate with which alternative causes occur, b_A , is not known in our example, because alternative causes remained unobserved. The causal powers that generate the observed data are represented by the weights, w_C and w_A , of the causal arrows that connect C and A to E . These parameters correspond to the probabilities with which the target cause and alternative causes generate E individually.

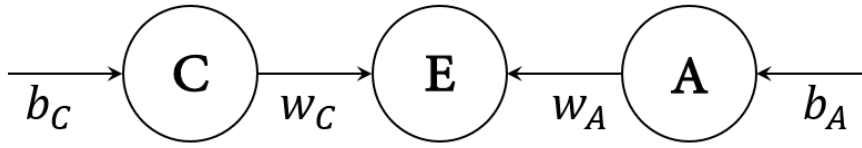


Figure 2.2: The common-effect causal structure assumed by the Causal Power Theory as the underlying generative causal source of the observed contingency. C and E in the causal structure denote the target cause and effect factor, respectively. A is assumed to comprise the sum of all known and unknown alternative causes of E . b_C and b_A denote the base rates of C and A ; w_C and w_A denote the causal powers of C and A . C and A are assumed to combine their causal powers according to a noisy logical OR function.

The validity of the causal structure depicted in Figure 2.2 as an explanation of the observed data rests on the following independence assumptions: First of all, C and A must occur with independent base rates, which means that the marginal probabilities with which C and A occur, $P(c)$ and $P(a)$, must remain invariant in the presence and absence of the respective other factor (e.g., $P(a|c) = P(a|\neg c) = P(a)$ and $P(c, a) = P(a) \cdot P(c)$). Note that this is a criterion whose fulfillment a professional researcher tries to accomplish in an experimental setting by means of manipulation of the target cause factor and random allocation of the objects (e.g., patients) of measurement to the different conditions. Secondly, C and A have to generate E with independent, non-interacting causal powers. This means that the probability with which C or A tend to generate the effect in isolation have to stay invariant in each others' presence. The third assumption that must be met is that the causal powers of C and A are independent of the base rates with which C and A occur, that is, a change of the base rate parameters b_C and b_A of the causal structure in Figure 2.2 must not alter the weights w_C and w_A of the causal arrows that lead into E .

With these assumptions at hand, we can see how the theory explains the two components that determine ΔP . First, let us consider how the theory explains the observed base rate of the effect, $P(e|\neg c)$, which is the second component of ΔP . According to the theory, all the effects that have occurred in the absence of C must be due to the isolate influence exerted by the alternative cause factors A . The right panel of Figure 2.3 shows what we can assume has happened to the causal structure depicted Figure 2.2 in the control group. As C is prevented from occurring through experimental intervention in this group, $P(e|\neg c)$ is given by

$$P(e|\neg c) = b_A \cdot w_A, \quad (2.2)$$

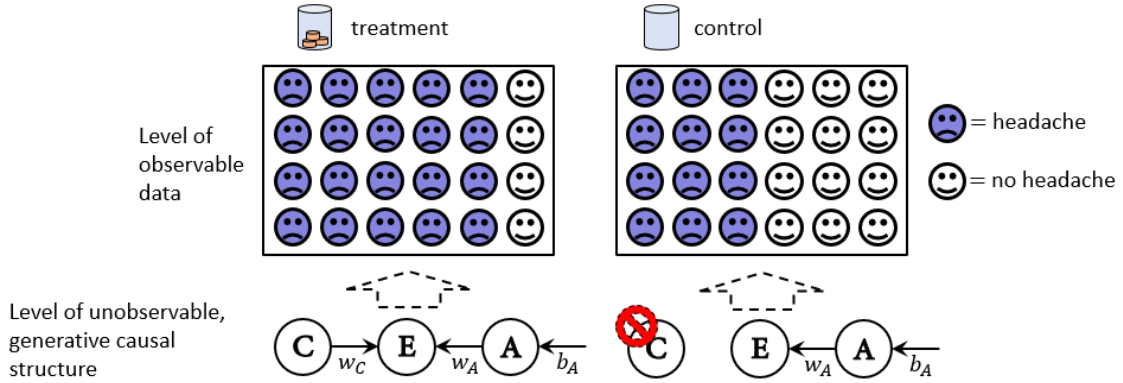


Figure 2.3: Illustration of how the Causal Power Theory explains the observed contingency by means of an underlying common-effect structure in which only the status of the target cause factor C is assumed to vary between the control and the treatment condition.

the product of the base rate with which alternative causes occur and their causal power. In the present example this product is known to be $b_A \cdot w_A = \frac{1}{2}$, as we have seen that half of the patients in the control group suffered from headaches.

By contrast, the probability of the effect in the presence of the cause, $P(e|c)$, which represents the first component of ΔP , is assumed to have resulted from a joint influence of the target cause and the alternative causes (see left panel in Figure 2.3). If the independence assumptions that were introduced above are met, that is, if $b_A \cdot w_A$ remains invariant in the presence of C , $P(e|c)$ is given by

$$P(e|c) = w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A, \quad (2.3)$$

the sum of the causal power of C and the product of the base rate and the causal power of the alternative causes A minus the overlap of the three parameters. This equation instantiates the assumption that the effect can either be caused by the target cause C or by the alternative cause A or by both simultaneously. With these two equations, the formula for ΔP can be rewritten as a function of the parameters of the causal structure depicted in Figure 2.2:

$$\Delta P = P(e|c) - P(e|\neg c) = w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A - b_A \cdot w_A. \quad (2.4)$$

Having derived this equation, we can now see how the causal power of C can be obtained. As the product of the base rate and the causal power of the alternative factors ($b_A \cdot w_A$) is assumed to be measured by $P(e|\neg c)$, the only quantity that remains unknown is the causal power of C , w_C , which we can see if we rewrite the

equation for ΔP as

$$\Delta P = w_C + P(e|\neg c) - w_C \cdot P(e|\neg c) - P(e|\neg c). \quad (2.5)$$

If we now rearrange this equation, we see that the generative causal power¹ w_C of the target cause factor C can be computed from the observed data with the following equation:

$$w_C = \frac{\Delta P}{1 - P(e|\neg c)}. \quad (2.6)$$

In the present example, applying Equation 2.6 reveals that the causal power of the hypothetical medicine is $w_C = \frac{2}{3}$, that is, the medicine tends to generate headaches with a two-thirds probability. In practical terms, two out of three headache-free patient swallowing the pill can expect to develop headaches.

The principle behind Equation 2.6 can also be illustrated by means of counterfactual considerations in the following way (cf. Pearl, 2000). If we assume that the alternative causes A have occurred with the same probability in the treatment group as they did in the control group, and if we also assume that they have exerted the same influence in the treatment group as they have done in the control group, we can conclude that half of the patients in the treatment group would have developed headaches even if they had not taken the medicine. In this case there would have remained twelve patients in the treatment group in which the medicine could have made an exclusive difference. If this had actually happened, we would have seen all patients suffering from headaches in the treatment group. However, the results show that four patients did not end up with headaches, which implies that out of the twelve hypothetical cases in which C could have made an exclusive difference, it actually seems to have done so in only eight. That is, the strength or causal power with which C tends to generate the effect is $\frac{8}{12} = \frac{2}{3}$.

Another possibility to illustrate the logic behind the Causal Power Theory is to use an Euler diagram (cf. Novick & Cheng, 2004) as shown in Figure 2.4, which I introduce here because it will be helpful in the following chapters. Like in Figure 2.1, the panel containing the observations that were made in the presence of C are shown on the left and the panel containing the observations that were made in the absence of C are shown on the right. All the effects that occurred are represented by the white area, whereas the shaded area depicts all cases in which the effect did not

¹Cheng (1997) also provides a set of assumptions and the equations for the case of preventive causal power, the probability with which a cause tends to prevent an effect from occurring. However, as the concept of generative causal power is the relevant one for this thesis, the equations for preventive causal power will be omitted here.

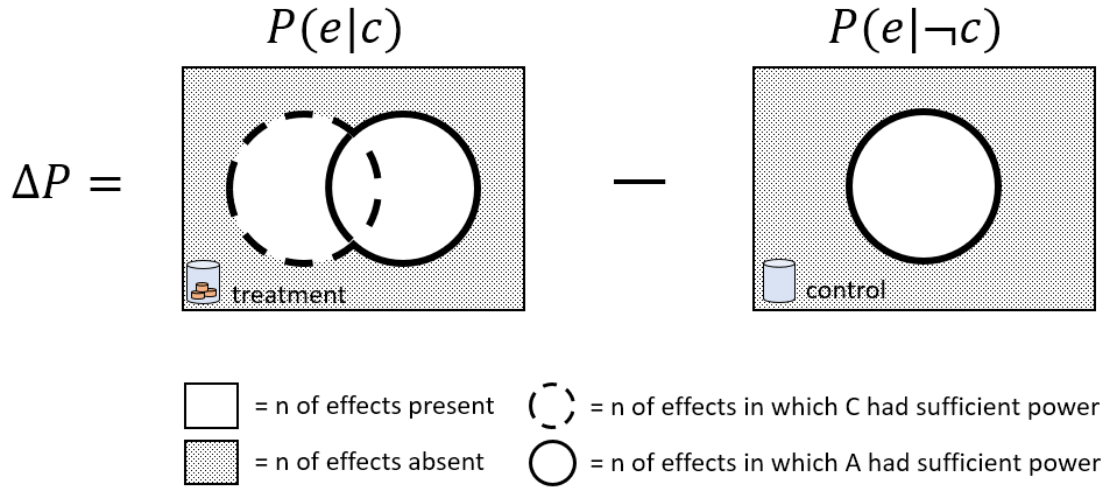


Figure 2.4: Euler diagram illustrating how the causal power theory explains the observable ΔP by means of unobservable causal powers exerted by two independent causal sources, C and A (cf. Novick & Cheng, 2004).

occur. The two different types of circles (dashed vs. solid) represent the two sources of generative causal power, C and A , that are assumed by the theory. The fact that no white space lies outside of the two circles illustrates that C and A together are assumed to comprise all possible causes of the effect, as well as the assumption that effects can never occur without being caused. Further, the assumption that the probability with which the background factors A occur is independent of C and the assumption that the causal power of A remains constant in the presence of C is captured by the constant size of the solid circle in the left and the right panel. The assumption that the causal power w_C of the target cause C adds independently and according to a noisy logical OR function to the influence of the background causes A (formally given by $w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A$ in Equation 2.3) is captured by the white space being larger in the left than in the right panel and by there being an overlap of the dashed and the solid circle. With the help of this diagram, we can see that Equation 2.6 estimates the individual causal influence of C by partialling out from the observable $P(e|c)$ the influence of the alternative causes, given by $P(e|\neg c)$.

2.2 Evidence for the psychological validity of the theory

The Causal Power Theory represents a normative computational theory (Anderson, 2013; Marr, 1982) of causal inference that aims at providing an optimal solution to the inference problem. However, the theory also has been shown to be descriptively

accurate in many contexts. Most importantly, the theory can capture several phenomena observed in human causal inference that purely associative accounts, like the Rescorla-Wagner Model (Rescorla & Wagner, 1972) or the ΔP Model, which do not distinguish between the data level and the level of an underlying generative causal source, fail to explain (comprehensive recent overviews are given, for example, by Cheng & Buehner, 2012; Cheng & Lu, 2017).

2.2.1 Zero-contingencies under ceiling vs. non-ceiling conditions

One phenomenon that the Causal Power Theory can readily explain and that purely covariational accounts that lack any causal assumptions that go beyond the data given cannot account for is why a probabilistic contrast of $\Delta P = 0$ is regarded as an indicator for the absence of a causal capacity of the target cause factor to generate the effect only if $P(e|c) = P(e|\neg c) < 1$ but not if $P(e|c) = P(e|\neg c) = 1$. Imagine our fictitious study had yielded the perfect ceiling effect that is illustrated in Figure 2.5 A). If all patients had suffered from headaches at the time of measurement, ΔP would have been zero ($P(e|c) - P(e|\neg c) = 1 - 1 = 0$). Yet, instead of having concluded that the medicine is generally ineffective, we rather would have been inclined to say that the causal status of the medicine cannot be determined under these conditions. Cheng’s Causal Power Theory captures this intuition, as Equation 2.6 is undefined in this case. Conceptually, if the medicine had not been administered, we would still have expected all patients in the treatment group to exhibit the effect due to the alternative causes. This, in turn, implies that there would not have been any chance or “space” for the medicine to display its potential causal power. The typical experimental strategy that a researcher would pursue in such a case would be to try to identify important alternative cause factors and to shield the experimental setting from their influence.

Now contrast this case with the situation that is depicted in Figure 2.5 B), in which ΔP is zero but $P(e|c)$ and $P(e|\neg c)$ are smaller than one. Imagine that these were the results the researcher obtained in her second attempt to test whether the medicine has an effect after she had managed to eliminate an important alternative cause factor of headaches. Here, the absence of a difference between the two groups intuitively seems to signal a lack of causal power of the candidate cause. In line with this intuition, Equation 2.6 is zero in this case. We can imagine that the medicine would have had the chance to make an exclusive difference in half of the treatment group, but the fact that this difference was not observed intuitively is taken as an indication that the medicine is lacking the hypothesized causal capacity.

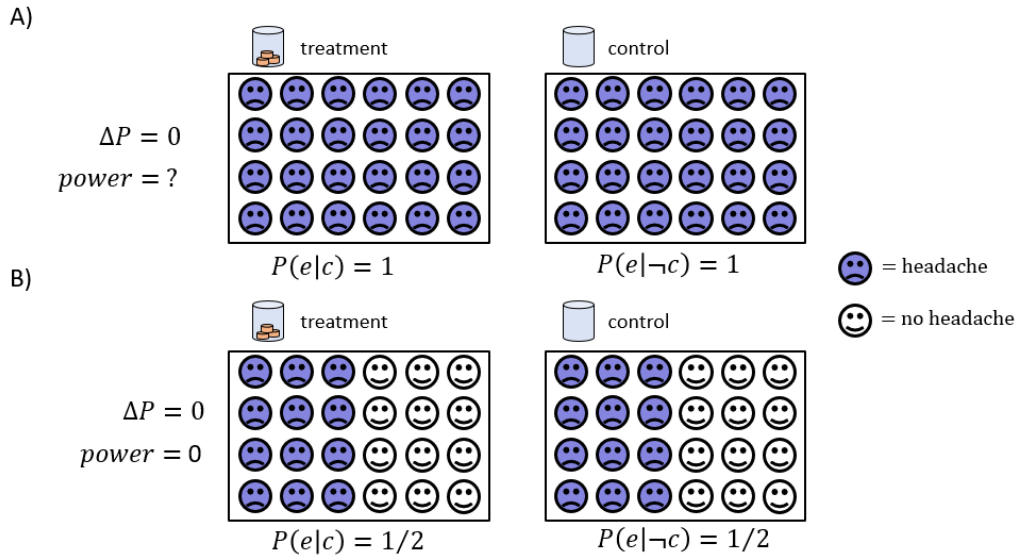


Figure 2.5: Different situations in which the contingency is zero but intuitions with respect to the causal power of target cause factor vary. A) represents a non-contingent situation due to a perfect ceiling effect. B) represents a non-ceiling situation in which ΔP is zero.

2.2.2 Different intuitions towards constant positive probabilistic contrasts

Another type of situation that provides evidence that lay people's causal inferences are driven by assumptions in line with the Causal Power Theory is one in which constant positive values of ΔP result from different combinations of $P(e|c)$ and $P(e|\neg c)$ (see Cheng, Novick, Liljeholm, & Ford, 2007). An example for such a case is shown in Figure 2.6. As can be seen, ΔP is one-third in both Situation A and Situation B. However, in Situation A this value of ΔP is obtained in a context in which the marginal probability of the effect, $P(e)$, is relatively high, while in Situation B the same value of ΔP is obtained under a lower level of $P(e)$. While these two situations are treated as equivalent by purely associative models, applying Equation 2.6 reveals that the Causal Power Theory predicts higher causal strength inferences in Situation A than in Situation B. In Situation A, the candidate cause appears to have made a difference in two-third of the cases that cannot be explained by the alternative causes, and hence its causal power is $w_C = \frac{2}{3}$. In Situation B, by contrast, the cause appears to have made a difference in only two-fifths of the cases that cannot be explained by the influence of the alternative causes, and hence its causal power is only $w_C = \frac{2}{5}$. Generally, given a constant positive probabilistic contrast, the Causal Power Theory predicts increasing values with an increase in the effect's base rate. An interesting aspect about the two data sets shown in Figure 2.6

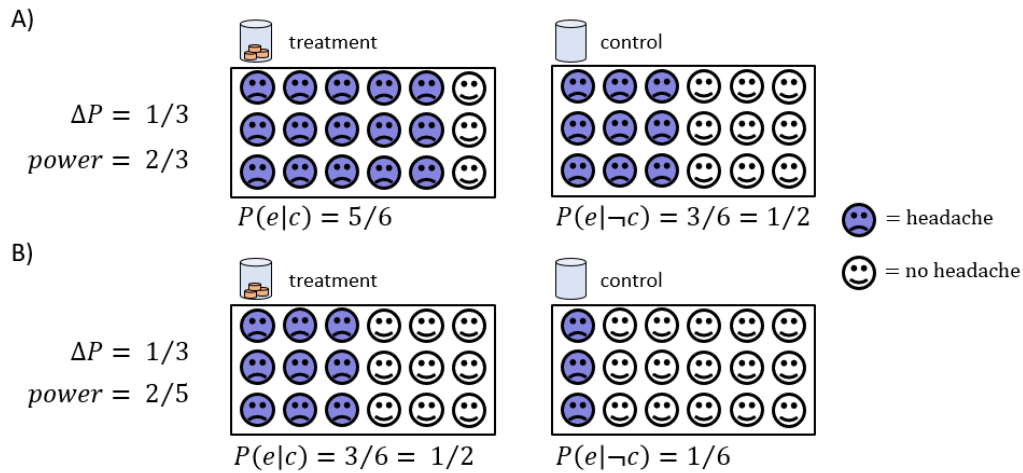


Figure 2.6: Different situations in which the contingency is constant at a positive value but intuitions with respect to the causal power of target cause factor vary.

is that a statistical analysis with a χ^2 test, which in contrast to the Causal Power Theory does not make any a priori causal assumptions, would yield identical p -values of 0.03.

Buehner et al. (2003) in their Experiment 2 confronted subjects with samples of observations that were analogous to those contrasted in Figure 2.6. After subjects had observed the respective data set, they were first asked to estimate whether they believed that the medicine has an effect or not. If subjects indicated an influence, they were further asked to estimate how many of one hundred fictitious new patients who would all not be suffering from headaches could be expected to develop headaches if they took the medicine. This question was supposed to elicit causal strengths ratings. In line with the predictions made by the Causal Power Theory, subjects in the condition in which the marginal probability of the effect was high gave higher causal strength ratings than subjects in the condition in which the same contingency was observed under a lower level of $P(e)$.

2.2.3 Non-interactive causal powers as a tacit default assumption

A further demonstration for the psychological validity of the core assumptions of the Causal Power Theory was provided by Liljeholm and Cheng (2007). In their experiments, Liljeholm and Cheng (2007) investigated whether lay people tacitly share the assumption made by the theory that the candidate cause and the alternative causes generate the effect with non-interactive causal powers. This assumption of the theory is particularly interesting because, other than the assumption that target

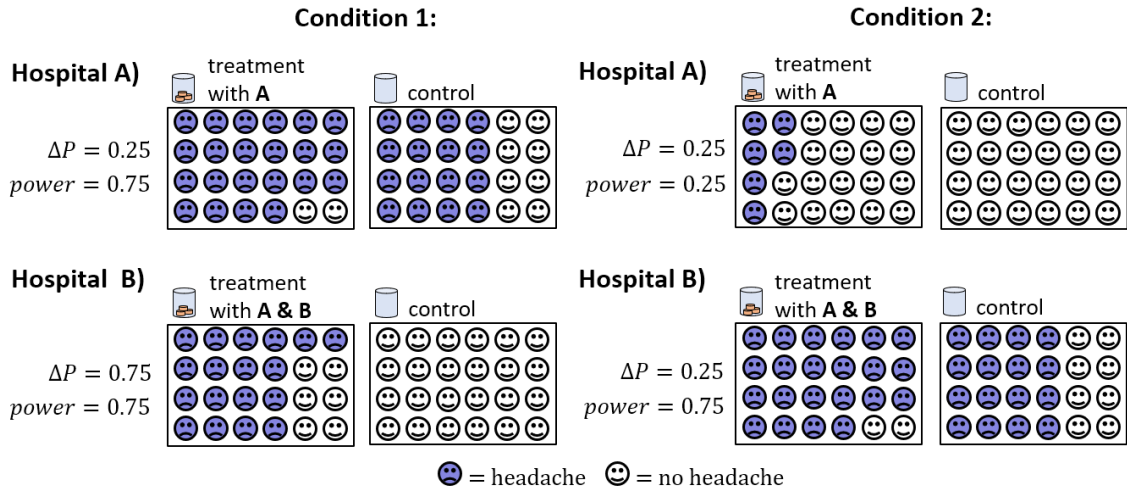


Figure 2.7: The two different contingency data sets that Liljeholm and Cheng (2007) presented their subjects in the two conditions of their study.

cause and alternative causes occur with independent base rates, its fulfillment lies beyond the control of a reasoner. The simplest method to establish independent base rates of the different potential causes of the target effect (i.e., to ensure non-confounding) is randomization and manipulation, the principles that characterize a scientific experiment. Nothing can be done, however, to prevent the candidate cause from interacting with present alternative causes of the effect. As Cheng and Lu (2017) have described it, the assumption of non-interactive causal powers therefore reflects a reasoner's aspiration that causal knowledge that has been acquired in the past remains useful for application in novel contexts that will likely be different with respect to the causal background factors. Of course, this assumption can be given up in light of contradictory empirical evidence (see Cheng & Lu, 2017, for an elaboration on this issue).

The data sets that Liljeholm and Cheng (2007) presented to their subjects in one of the experiments are depicted in Figure 2.7. Subjects read that a pharmaceutical company wanted to investigate whether two different types of an allergic medicine (Medicine *A* and Medicine *B*) might cause headaches as a side effect, and that the company had run two different studies to investigate this. Subjects then were either presented with the two contingency data sets shown in the left panel of Figure 2.7 (Condition 1) or with the two contingency data sets shown in the right panel of Figure 2.7 (Condition 2). Subjects were asked to assume that the two studies were carried out in two different hospitals (Hospital *A* and Hospital *B*) and that the probability with which headaches occur at base line might differ. While Hospital *A* tested the isolate influence of medicine *A*, Hospital *B* tested the influence of Medicine *A* and Medicine *B* in combination. Crucially, the causal test question that subjects

were asked after they had inspected the results of the fictitious studies referred only to Medicine *B*, the medicine that had never been seen being tested in isolation. Subjects were asked to indicate whether they believed that Medicine *B* has the power to cause headaches as a side effect.

What Liljeholm and Cheng (2007) wanted to find out with this test question was which assumptions about the causal properties of Medicine *A* their subjects had made when they evaluated the data set showing the effects of the combined treatment. The hypothesis was that, if subjects tacitly made the default assumption that the causal power of Medicine *A* remained invariant in the context where Medicine *B* was simultaneously present, subjects in Condition 1 should have come to the conclusion that Medicine *B* is causally ineffective, while subjects in Condition 2 should have inferred that Medicine *B* has the causal capacity to generate the effect. To see this, let us first consider Condition 1. From the first panel showing the results obtained in Hospital *A*, we can see that Medicine *A* in isolation tends to generate headaches with a causal power of 0.75: it can be expected that two-thirds of the patients in the treatment group would have developed headaches due to the background causes even if Medicine *A* had not been administered. Thus, in the remaining one-third of the patients in which only Medicine *A* could have made a difference, it actually seems to have done so in 75%. Now consider the results obtained in Hospital *B*. First of all, the observations made in the control group indicate that alternative background causes were, other than in Hospital *A*, apparently not an issue in Hospital *B*. Now, if we assume that Medicine *A* exerted the same causal power in the presence of Medicine *B* as it did in Medicine *B*'s absence (in the context of Hospital *A*), then we should see more than 75% percent of the patients exhibiting the effect only if Medicine *B* has been causally effective, too. What we can see, however, is that from the twenty-four patients who took the combination of drugs, only eighteen ended up with headaches, that is 75%. Medicine *B* therefore seems to be ineffective.

Now consider what happened in Condition 2. Here, Medicine *A* produces the effect with a power of 0.25. Now, if Medicine *B* was ineffective and Medicine *A* exerted its power invariantly, only two out of the eight patients that can be assumed not to have developed headaches in the absence of any treatment should have been affected. We can see, however, that six out of these eight cases developed headaches, and therefore can conclude that Medicine *B* possesses a causal capacity to bring about headaches. (In fact, Medicine *B* is even stronger than Medicine *A* according to the Causal Power Theory; its causal power is $\frac{2}{3}$). Another way to see this is to apply Equation 2.6 first to the results obtained in Hospital *A* and then to the results

obtained in Hospital *B* by simply ignoring that Medicine *B* was also present there. Equation 2.6 would yield a causal power of *A* of 0.75 in the context of Hospital *B*, which is a strong deviation from what is obtained if the Equation 2.6 is applied to the data from Hospital *A*. This deviation can be explained by an additional independent influence of Medicine *B*.

The vast majority of subjects in both conditions provided answers that were in line with the predictions made by the Causal Power Theory. These results indicate that reasoners indeed seem to try to estimate the power of causes when they evaluate contingency information and that they tacitly make the default assumption that causal powers remain invariant across contexts that change with respect to the composition of background factors. Importantly, these results are again inexplicable under a purely covariational view lacking any a priori causal assumptions that constrain the interpretation of the learning data (see also Cheng, Liljehom, & Sandhofer, 2013).

2.3 Problematic findings

A crucial prediction of the Causal Power Theory that was empirically confirmed by the study of Buehner and May (2003) is that judgments of causal strength should, under constant *positive* levels of ΔP , increase with the observed base rate of the effect, $P(e|\neg c)$. This positive influence of $P(e|\neg c)$ should, however, not occur under conditions of $\Delta P = 0$. Yet, the results of Experiment 1 in Buehner et al. (2003) revealed a similar positive influence of $P(e|\neg c)$ under these conditions, an effect that has been called “frequency illusion” in the literature (e.g., Allan & Jenkins, 1983).

Other problematic findings have been reported by Lober and Shanks (2000), who tested data sets in which causal power was kept constant, while ΔP was varied. For example, in some of the data sets that Lober and Shanks (2000) showed their subjects the effect always occurred in the presence of the cause, $P(e|c) = \frac{28}{28}$, while the base rate decreased in steps of 0.25 from $P(e|\neg c) = \frac{21}{28}$ over $P(e|\neg c) = \frac{14}{28}$, and $P(e|\neg c) = \frac{7}{28}$ to $P(e|\neg c) = \frac{0}{28}$. The causal power of the target cause is 1 in all these cases. Yet, subjects’ ratings increased with a decrease of $P(e|\neg c)$.

The problematic findings of Buehner et al. (2003) and Lober and Shanks (2000) have been taken as evidence by Griffiths and Tenenbaum (2005) for their Bayesian Causal Support Model, according to which reasoners are sensitive to and express in their causal judgments the degree to which an observed contingency “supports” the inference that the target cause does or does not possess the causal power to generate the effect. Griffiths and Tenenbaum (2005) have pointed out that causal power rep-

resents a mere *point estimate* of the strength of a causal connection, which neglects the possibility that any empirical effect (the observed probabilistic contrast) could have resulted from mere *sampling variation*. With respect to the causal diagram shown in Figure 2.2, while causal power assesses the strength of a causal relationship (the weight of the causal arrow, w_C), the question whether an observed effect is due to mere sampling variation or due to an actual influence of the target cause refers to the existence of the causal arrow connecting C and E . For an illustration that intuitions about the reliability of the observed empirical effect plays a crucial role, contrast the two data sets tested in Lober and Shanks (2000) in which $P(e|\neg c)$ was either $\frac{21}{28}$ or $\frac{0}{28}$, while $P(e|c)$ was $\frac{28}{28}$. In both cases, the point estimate for causal power is $w_C = 1$, but the confidence with which we can conclude that C possesses the power to produce the effect and that it does so as strongly as suggested by the data seems to be higher in the latter case, where $P(e|\neg c) = 0$. It is very unlikely in this latter case but not too implausible in the former case that all observed effects in the presence of C were actually caused by A if the effect was never observed in C 's absence. Interestingly, the ratings that subjects made in the $P(e|\neg c) = \frac{21}{28}$ condition were indeed close to 0.5, in line with the hypothesis that subjects were uncertain as to whether the data provided evidence for a positive effect. The observed positive influence of $P(e|\neg c)$ under zero-contingencies is explained in the same way. Zero-contingencies provide better evidence for a causal power of $w_C = 0$ if $P(e|\neg c)$ is low (see also Liljeholm & Cheng, 2009, who explain these effects with variations in the virtual sample size, which is the number of entities in the C -present sample in which the target cause factor can reveal its efficacy unambiguously). According to the analysis of Griffiths and Tenenbaum (2005), the “frequency illusion” thus reflects a normative reasoning strategy.

The problem that the Causal Power Theory provides only point estimates for the strength of a causal relationship and neglects the possibility that an observed relationship might have been observed as a result of sampling variation will be addressed again in Chapter 5, where I discuss the implications of this problem for the assessment of singular causation.

2.4 Summary

This chapter has introduced Cheng’s (1997) Causal Power Theory, together with some core phenomena that the theory explains elegantly and phenomena that are problematic for the theory. The core idea of the theory is that reasoners explain observable events in virtue of unobservable causal powers that made these events

happen. Moreover, it is assumed that it is the operation of unobservable causal powers that gives rise to observable patterns of covariation, and that observable patterns of covariation between a potential cause and effect factor can, under particular conditions, be used by reasoners as a diagnostic tool to infer the causal power of a candidate cause factor. The causal power or strength of a candidate cause can be operationalized as the probability with which it makes its effect happen.

Chapter 3

Applying Causal Power Knowledge to Evaluate Singular Cases – The Power Framework of Causal Attribution

“How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?”

- Arthur Conan Doyle, stated by Sherlock Holmes in *The Sign of Four*

We have seen how a reasoner can learn the causal power with which a target cause tends to generate its effect from a set of observations that track the covariation between a target cause and target effect factor. In the present chapter, Cheng and Novick’s (2005) Power Framework of Causal Attribution will be introduced, which represents a computational account telling us how this knowledge about the power of candidate causes ought to be applied to determine causality in singular cases in which it is known that the effect has actually occurred. We will first see how the framework works and which principles that seem to be important for the assessment of singular causation it is acknowledging. We will then turn to what I propose is its *conceptual* weak spot, the incapability to account for cases of causal preemption.

For a better illustration, it will be helpful to have again a concrete example. Let us imagine we get to know a particular person, Suzy. What we know about Suzy is that she was a volunteer in our fictitious study (the results of which are depicted in Figure 2.1) in which it was investigated whether the newly developed medicine can cause headaches as a side effect. What we also know about Suzy is that she is suffering from headaches, and we now wonder whether her episode of headaches has occurred because she took the new medicine. In particular, what we want to know

in this case seems to be: what is the probability, or, to use the Bayesian reading of probability, the degree to which we can believe, that her suffering from headaches is due to the medicine and not due to something else. What we have learned already is that the medicine has a causal power of two-thirds, which means that someone not suffering from headaches who takes the medicine can be expected to end up with headaches with a two-thirds probability. According to the Power Framework of Causal Attribution, this is crucial information to assess the probability with which the observed effect is due to the target cause. However, the causal power of the cause is obviously not the target quantity here, because we already know that Suzy is suffering from headaches. The situation is, to use a standard terminology from probability theory, conditionalized on the presence of the effect. Our goal is to see how well our target cause serves as a causal explanation for *this* effect on *this* occasion. Let us see how the Power Framework of Causal Attribution attempts to handle this case.

3.1 Different equations for different states of knowledge about the target cause

Our point of departure is a situation in which we know that the target effect factor (E) had been instantiated (e) and the question we want to answer is how likely it is, or, to use a Bayesian terminology of probability, how strongly we can believe that a potential target cause factor (C) was actually causally responsible for the observed occurrence of the effect. Cheng and Novick (2005) have proposed different equations aiming to compute this probability, which differ with respect to the additional information about the status of the potential cause that a reasoner has. For example, thus far all we know about Suzy is that she is suffering from headaches, whereas we do not know whether she had been a participant who was allocated to the control group or to the treatment group. In other words, we do not know whether the potential cause we have under suspicion had actually been instantiated in Suzy or not. According to the Power Framework of Causal Attribution, the probability that Suzy's episode of headaches is due to the medicine in this situation can be formally written as $P(c \rightarrow e|e)$, where the arrow that points from c to e represents a singular or *actual* causal connection between the two events. This probability is supposed to be given by the following equation:

$$P(c \rightarrow e|e) = \frac{b_C \cdot w_C}{b_C \cdot w_C + b_A \cdot w_A - b_C \cdot w_C \cdot b_A \cdot w_A} = \frac{b_C \cdot w_C}{P(e)}. \quad (3.1)$$

Equation 3.1 shows that what we should do according to the framework in this situation is to normalize the product of the base rate with which the target cause occurs and its causal power by the marginal probability with which the effect occurs. In the present example, we know that the base rate of taking the medicine is 0.5, as half of the participants had been allocated to the treatment group and half to the control group. Checking once again the results we can further see that the effect occurred in thirty-two of the forty-eight subjects who took part in the study, which yields a marginal probability of the effect of $P(e) = \frac{2}{3}$. We can hence calculate that the probability that Suzy's episode of headaches was actually caused by the medicine is $P(c \rightarrow e|e) = \frac{\frac{1}{2} \cdot \frac{2}{3}}{\frac{2}{3}} = \frac{1}{2}$. Under a Bayesian interpretation of probability, we should be uncertain as to whether it had actually been the drug that generated Suzy's headaches.

Now imagine that we asked the head of the study to check and tell us whether Suzy was part of the treatment or the control group and that the answer we get was that Suzy had been part of the former. In this case, we could conditionalize on the presence of the target cause factor and Equation 3.1 would turn into

$$P(c \rightarrow e|c, e) = \frac{w_C}{w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A} = \frac{w_C}{P(e|c)}, \quad (3.2)$$

the causal power of the medicine being normalized by the probability of headaches that occur when the medication is taken. In this case the probability that the medicine actually caused Suzy's headache is supposed to increase from $P(c \rightarrow e|e) = \frac{1}{2}$ to $P(c \rightarrow e|c, e) = \frac{4}{5}$. That is, we should now conclude that her headaches have resulted with a probability of eighty percent due to the medicine.

Equation 3.2 will be the relevant one for the subsequent conceptual analyses. First of all, focusing on occasions on which it is known that the target cause factor was instantiated allows us to ignore the base rate parameter of the target cause, which simplifies the calculations. The second reason why we can focus on Equation 3.2 is that the problematic implication of the framework does not seem to be that the probability with which a target cause is actually instantiated in the target case should be a relevant piece of information. The noteworthy and, as will turn out, problematic implication of the framework is that it predicts that we should say that a potential cause event c was the singular cause of an observed effect e whenever c has been sufficient to produce the effect on the occasion.

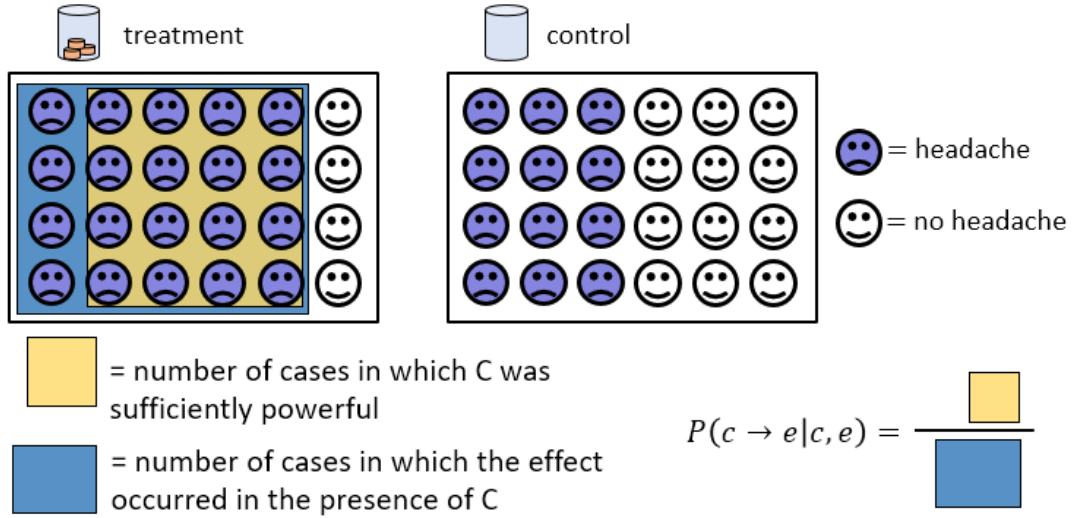


Figure 3.1: Illustration of the principle behind Equation 3.2. The yellow area represents the cases in which the target cause was probabilistically sufficient to generate the effect, which are formally given by $w_C \cdot [n(c, e) + n(c, -e)]$. The blue area marks all effects that occurred when C was instantiated, formally given by $P(e|c) \cdot [n(c, e) + n(c, -e)]$. Equation 3.2 can be understood as computing the ratio of the yellow and the blue area.

3.2 A filter of probabilistic sufficiency

What Equation 3.2 essentially represents is a “filter of probabilistic sufficiency”, as it identifies among all observed co-instantiations of C and E the relative frequency of cases in which C was sufficiently powerful to generate the effect. Figure 3.1 illustrates the principle behind Equation 3.2 by means of the absolute numbers that were depicted in the results graphic of our fictitious study. The blue area in the left panel of Figure 3.1 marks all the effects that occurred when the target cause was present. As our analysis focuses only on those occasions on which C was actually instantiated, we can ignore the right panel showing what happened in the control group. We have seen earlier that the predictive probability was $P(e|c) = \frac{5}{6}$, and so the blue area comprises the twenty out of the twenty four subjects in the treatment group who showed the effect. We also have learned that the medicine has the capacity to produce headaches with a two-thirds probability. Therefore, we know that the medicine was probabilistically sufficient to generate headaches in sixteen out of the twenty-four treatment-group subjects. These cases are marked by the yellow area in the Figure 3.1. If we now divide the number of cases contained in the yellow area through the number of cases contained in the blue area, we obtain $\frac{16}{20} = \frac{4}{5}$, the quantity computed by Equation 3.2.

Another way to illustrate the output that Equation 3.2 delivers is the Euler

diagram that we have seen earlier in Figure 2.4. Remember that all cases in which the target cause was sufficiently powerful to generate the effect were marked there by the dashed circle and that all the effects that occurred in the treatment group corresponded to the white area within the two circles. Equation 3.2 can therefore alternatively be understood as the ratio of the dotted circle's area and the area contained in the two circles.

3.3 What the framework seems to get right

What we can see from Equation 3.2 is that the assessment of singular causation should focus not on the absolute but on the relative causal power of the target cause in comparison to the prevalence and the causal power with which alternative causes bring about the effect. This principle of the framework is in line with two sorts of intuitions that enter the assessment of singular causation.

3.3.1 The stronger the target cause the better

One intuition that Equation 3.2 reflects is that our degree of belief in a singular causal connection between c and e should increase if the probability with which the target cause generates the effect increases relative to the prevalence and the power of the background causes. Just imagine that the base rate remained as was observed in our fictitious study but that the causal power of the medicine was not two-thirds but, for example, five-sixths. (In this case, we would have seen twenty-two instead of twenty patients suffering from headaches in the treatment group. In this case Equation 3.2 would have yielded a value of 0.91.) It appears to be intuitively right to say that we should be more convinced that c caused e on a singular occasion observed under these conditions than under the former conditions.

3.3.2 The less likely and the weaker alternative causes the better – Holmesian inference

Another principle that is captured by Equation 3.2 is that our degree of belief in c having caused e on an occasion should, everything else being equal, become stronger the lower the probability is with which alternative causes are present and the weaker they are. This principle incorporated by the framework captures the notion that is, for example, inherent in what the philosopher Alexander Bird (2005; 2007; 2010) calls “Holmesian inference”, the reasoning process by which one of multiple potential

and mutually exclusive explanations for a phenomenon is selected based on the relative probabilities of the competing explanations (see also Lipton, 2004; Lombrozo & Vasilyeva, 2017, for related accounts on abduction or *inference to the best explanation*). The dictum of Holmesian inference is captured by the quote at the beginning of this chapter. Think of how the extreme case that is captured by this quote relates to our scenario. Imagine that the base rate of the effect had been zero in our fictitious study, which would indicate the absence of sufficiently powerful alternative causes, and imagine that the medicine had still exerted its power as before. In this case, Equation 3.2 would have delivered a value of one (and so would Equation 3.1), in accordance with the Holmesian dictum. In philosophical terms, the model would have concluded in this type of situation that c must have been probabilistically sufficient to produce e on a singular occasion because c can be assumed to be *necessary* for e to occur. What we would see in our Euler diagram in such a case would be only the dashed circle in the left panel, illustrating that $P(e|c)$, the denominator of Equation 3.2, would be fully determined by the influence exerted by target cause factor. What the Euler diagram would also reveal is that the size of the causal power of the target cause becomes irrelevant in this case. The dashed circle could have any circumference and yet the output of Equation 3.2 would be unaffected.

3.4 The blind spot of the framework – the possibility of causal preemption

Now that we have seen how the framework works and which aspects about the assessment of singular causation it seems to be getting right, let us turn to what I claim is missing in the picture – sensitivity to the possibility that the target cause might have been preempted in its efficacy on an occasion by a potential alternative cause.

Scenarios of preemption refer to situations in which at least two potential causes are simultaneously sufficient to produce the effect but in which only one seems to be the singular cause of the effect. These cases are described in philosophical literature on causation (see, e.g., Danks, 2017; Hall, 2004; Paul, 2009; Paul & Hall, 2013; Strevens, 2008, for overviews) under the topic of *redundant causation*, and they have most intensely been discussed among philosophers who advocate a counterfactual view of causality (see, e.g., Halpern, 2016; Halpern & Hitchcock, 2015; Hitchcock, 2001, 2007, 2009; Lewis, 1973). According to a simple counterfactual theory of causality, a singular event c is the cause of another singular event e if it is true that e would not have occurred if c had not occurred. Cases of preemption threaten this

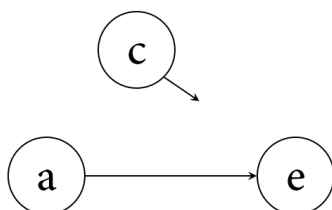


Figure 3.2: Neuron diagram modeling a situation of causal preemption (cf. Paul & Hall, 2013). The interrupted causal arrow departing from c and the arrow connecting a and e illustrate a situation in which c was causally preempted by a .

simple counterfactual dependence criterion of causation, whereas they appear to be less problematic (see, however, Paul & Hall, 2013) for *process* or *transference-based* accounts of causality (e.g., Dowe, 2000; Fair, 1979; Salmon, 1984). According to the transference-based view of causality, causation involves the transfer of quantity from cause to effect, where quantity, according to some accounts (e.g., Fair, 1979) can be physical energy or, according to other accounts (e.g., Dowe, 2000), any type of conserved quantity, such as linear momentum or charge (for an overview see, e.g., Dowe, 1995, or Waldmann & Mayrhofer, 2016).

One of the standard cases of causal preemption from the literature, in fact a case belonging to a type of preemption called *late preemption*, goes something like this (cf. Paul & Hall, 2013):

Two kids, Suzy and Billy, are known to be perfectly accurate rock throwers. Whenever either of them decides to fling a rock towards a bottle, the bottle shatters into pieces. On a particular occasion, it so happens that Suzy and Billy are aiming at the same bottle and both throw their stones with identical speed. Suzy, however, manages to throw her rock a split second earlier than Billy. The moment Billy’s stone reaches the location of the bottle it meets nothing but thin air.

The situation described by the scenario can be illustrated schematically with a neuron diagram as shown in Fig. 3.2 (see Paul & Hall, 2013). Node a represents the event “Suzy having thrown her stone” and node c represents the event “Billy having thrown his stone”. The arrow departing from a and reaching e is supposed to indicate that a was the singular cause of the effect. The arrow leaving from c but not reaching e illustrates that c had been preempted in its efficacy by a .

A doctrine among philosophers who aim to work out a definition of whatever type of concept, e.g., a reductive concept of causality, is to test whether their accounts survive the verdict that *common-sense intuition* provides for a set of key situations. As the philosopher David Lewis (1986) put it:

When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble.
(p. 194)

Scenarios of causal preemption are regarded as one such key type of case. Everybody who is presented with the scenario above has the intuition that Suzy’s rock throwing was the singular cause of the bottle’s shattering, whereas Billy’s rock throwing is considered as a *genuine non-cause*. A simple counterfactual account of causality, however, fails to identify Suzy’s rock throwing as the singular cause, for a simple counterfactual evaluation would reveal that the bottle would still have shattered if Suzy had not thrown her stone (for a more elaborate account that can handle this as well as other problematic cases see Halpern & Pearl, 2005).

We have seen earlier that Equation 3.2 represents a “sufficiency filter”, for it attributes causality to the target cause whenever the target cause is assumed to be probabilistically sufficient on an occasion. The Power Framework of Causal Attribution therefore has the opposite problem with situations of causal preemption. While it would identify Suzy’s rock throwing as the singular cause of the effect in the given example, it would fail to identify Billy’s throwing as a *non-cause* of the effect. The framework tends to identify events as singularly caused by c that actually seem to be solely due to an alternative cause a . That is, the model is too permissive when it comes to the assessment of singular causation. The failure to detect this asymmetry between the potential causes is a result of the framework’s focusing solely on static information about the relative causal powers of the potential cause factors, while it neglects that *temporal information* also seems to be crucial. The types of temporal information that are relevant will be introduced in the next chapter.

Importantly, the possibility of causal preemption does not only occur when the potential cause factors operate with deterministic causal powers, as was the case in the illustrative rock-throwing scenario. Rather, the possibility of preemption occurs whenever two potential cause factors are sufficiently powerful on an occasion. We have seen in our example about the fictitious new medicine that such occasions are assumed by the framework to arise also in the context of probabilistic causal powers. In the Euler diagram that was introduced, the cases in which the target cause and the alternative cause factors were simultaneously probabilistically sufficient for the effect were those cases enclosed by the overlap of the two circles. My proposal is that, if we had a “causal microscope” and could zoom into this overlap, we would find that it consists of three different types of cases¹: cases in which the target

¹It is important to note here that the framework considers the cause and effect factors to be

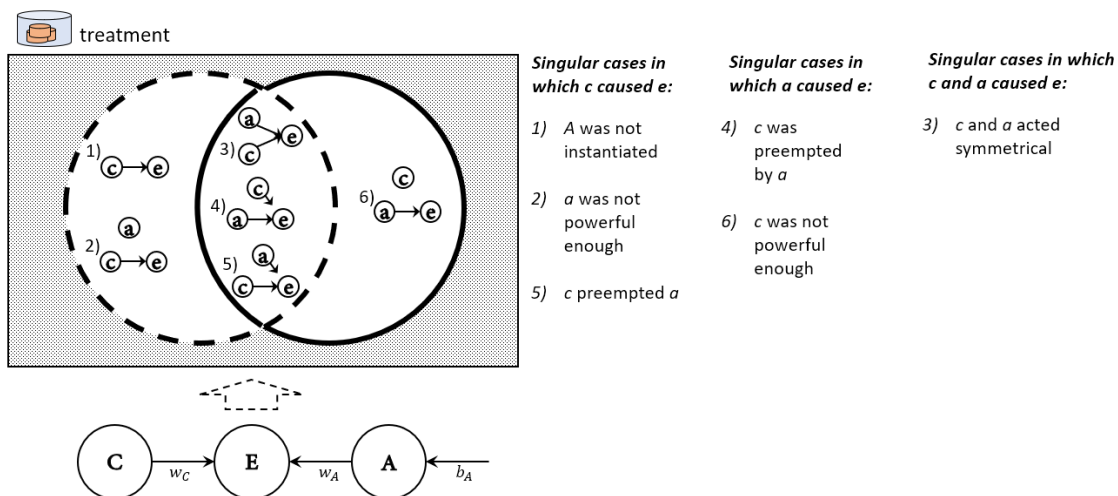


Figure 3.3: The singular causal structures (middle panel) that are compatible with (1) the general causal structure (bottom panel) assumed by the power framework and (2) with a singular observation (top panel) in which c and e have co-occurred.

cause preempted the potential alternative cause, cases in which the alternative cause preempted the target cause, or cases in which the target and the alternative cause exerted their causal powers symmetrically, a type of redundant causation that in the philosophical literature is called “symmetric overdetermination”. Whether two causes ever exert their powers perfectly symmetrical or whether we would find slight deviations if we zoomed in further is debatable, however. The different singular causal structures that are compatible with the general causal structure and with our example singular case Suzy are shown in Figure 3.3. The singular cases that correctly enter into $P(c \rightarrow e|c, e)$ have Singular Structures 1, 2, 3, and 5. The cases that are correctly excluded by Equation 3.2 are those exhibiting Singular Structure 6. The problem of Equation 3.2 is that it cannot distinguish between the three different singular structures of which the sufficiency overlap can be composed. By failing to exclude situations in which c was preempted by a (captured by Structure 4 in the Figure) from the estimation of $P(c \rightarrow e|c, e)$, the model tends to overestimate the probability with which c and e were actually causally connected on a singular occasion.

The degree to which Equation 3.2 overestimates $P(c \rightarrow e|c, e)$ depends on two factors: the size of the sufficiency overlap between C and A and the prevalence of situations in which c was preempted by a within the sufficiency overlap in a

binary variables that can either be present or absent. The factors involved in the medicine-headache scenario could also be plausibly represented on a continuous scale. In this case it would probably be more natural to think of an *additive* rather than a *logical-OR* combination of the potential cause factors, and intuitions regarding the possibility of causal preemption would probably be different.

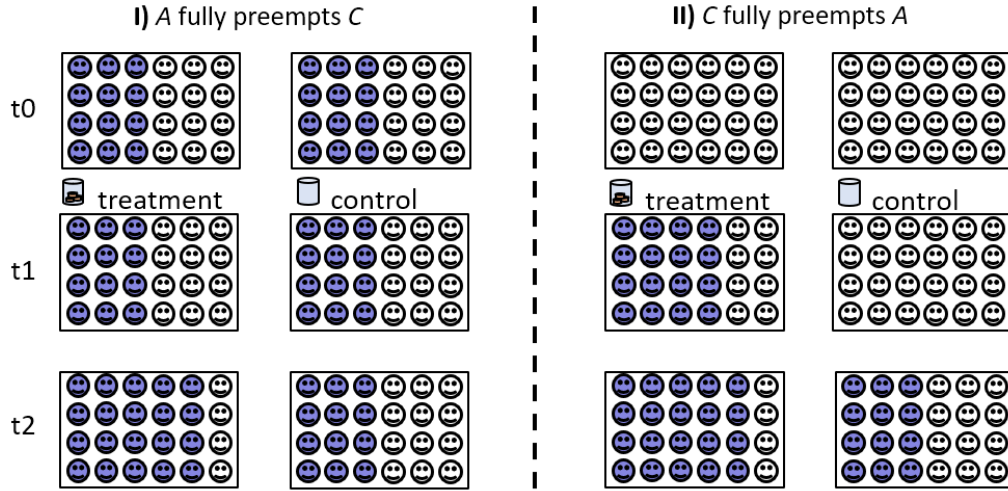


Figure 3.4: Illustration showing that the preemptive relation between target cause and alternative causes is conceptually unproblematic for the estimation of causal power.

particular context. Consider again the results of our fictitious medical study. We have seen there that the alternative causes produced the effect with a fifty percent probability in the control group. Accepting the independence assumption of the theory, we have seen that we should conclude that the alternative cause factors also were sufficiently powerful enough in fifty percent of the cases in the treatment group. This implies that there is a sufficiency overlap of $\frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$, that is, it can be expected that there are eight patients in the treatment group in which the medicine and the alternative factors were simultaneously sufficient to produce the effect. These cases are all attributed to the target cause by Equation 3.2, which implies that we should conclude that the medicine caused headaches in a singular case with a probability of eighty percent. However, in the extreme case in which the alternative factors fully preempted the target cause in its efficacy, the probabilistic overlap of $\frac{1}{3}$ should be attributed away from the target cause. $P(c \rightarrow e|c, e)$ would be cut in half and reduce to 0.40 in this case.

3.4.1 Preemption and the assessment of causal power

It was argued above that the possibility of causal preemption poses a problem for the assessment of singular occasion. It is important to note, however, that the preemptive relation between the potential causes of the effect does not, at least not theoretically, threaten the assessment of the general causal power of the target cause factor (Equation 2.6). As long as the assumption holds that C and A occur with independent base rates and as long as it is true that the two cause factors generate

the effect with independent causal powers, the preemptive relation between C and A does not affect the output delivered by Equation 2.6. Figure 3.4 is supposed to illustrate this by means of the two extreme scenarios of causal preemption. The left panel depicts the theoretically possible situation in which the alternative causes A produced their effects with $b_A \cdot w_A = 0.50$ at time t_0 , that is, even before the clinical intervention was carried out at time t_1 , whereas the causal power of the medicine did not become effective until t_2 . This scenario represents a type of preemption called “early preemption” in the literature (cf. Hitchcock, 2007), because one cause factor produced its effect even before the other cause occurred. The right panel depicts the situation in which the medicine exerted its power at t_1 before any alternative causes took effect. The left scenario is one in which the application of Equation 3.2 leads to a maximal overestimation of $P(c \rightarrow e|c, e)$, whereas the other extreme scenario on the right is the one in which Equation 3.2 seems to be correct. What can be seen is that the “endstates” at t_2 , however, are exactly identical in the two situations, which shows that the estimation of the causal power of the target cause obtained by applying Equation 2.6 is unaffected by the different cause factors’ preemptive relation².

3.5 Summary

This chapter has introduced Cheng and Novick’s (2005) Power Framework of Causal Attribution, which shows how knowledge about the strength of the potential causes, operationalized as causal power (Cheng, 1997), ought to be applied to assess causation in singular cases. It was shown that the Equations of the framework consider the relative causal power of the target cause and of the potential alternative causes to estimate the probability with which a particular target cause factor C was the singular cause of an observed effect E . This probability can also be understood in the Bayesian sense as the degree to which a reasoner can believe that a potential cause event c has caused the observed effect event e on an occasion. We have seen that the model captures the intuition that a target cause that brings about its effect relatively reliably should elicit a higher degree of belief that it was causal on a singular occasion than a target cause that is known to work relatively unreliably. It was also shown that the model incorporates the normative principle of abduc-

²While the principal validity of the causal power equation is not threatened by the preemptive relation between C and A , Figure 3.4 reveals that it may pose a practical problem, however. Figure 3.4 reveals that causal power should not be assessed before a causal end state has been reached in a context. A problem seems to be how, without a priori knowledge, one can know at what time this will be the case.

tive explanation, as it predicts that one can be certain that a particular factor was the singular cause of the observed effect, no matter how strong or weak this cause generally tends to bring about the effect, if alternative causes can be ruled out on an occasion. The problem of the framework, I have argued, is that its equations equate probabilistic sufficiency with causality, and as a consequence fail to account for the possibility of causal preemption. Causal preemption, mostly debated among *Causal Counterfactualists*, refers to types of situations in which multiple potential causes are simultaneously sufficient for the effect, but only one of them is considered to be the singular cause. Scenarios of preemption reveal that more is needed than knowledge about the potential causes' causal powers to be certain that an effect was singularly caused by the cause one has under suspicion. In the following chapter I will propose a generalization of Cheng and Novick's framework that is capable to handle the problem of causal preemption. The core idea of my proposal will be that knowledge about the temporal relation between the potential causes needs to be incorporated and combined with causal power information. The combination of causal power and temporal information enables us, so to say, to obtain an inferential version of "the causal microscope" that was mentioned above, with which we can disentangle what happens in the causal sufficiency overlap.

Chapter 4

A Generalization of the Framework Sensitive to Causal Preemption

“Time is what keeps everything from happening at once.”

- Ray Cummings *The Girl in the Golden Atom*

I argued that when a target cause and an alternative cause of an observed effect were simultaneously powerful enough to produce the effect on an occasion it needs to be decided whether the target cause was preempted in its efficacy by the alternative cause or not. If it was preempted, it should not count as the singular cause of the effect. We have seen that the equations of the standard Power Framework of Causal Attribution cannot handle this problem because they focus solely on the relative causal power of the target cause. The standard model seems to make correct predictions only in contexts in which the probability with which alternative causes preempt the target cause is zero. In the present chapter it will be argued that, to account for the possibility of causal preemption, what needs to be incorporated by the equations in addition to information about the powers of the potential causes is information about their *temporal relations*. I will present a refined model in which causal power and temporal information are combined to compute the probability with which the target cause was preempted by an alternative cause. I will first introduce the proposed refined equation and then elaborate on the different temporal components that are supposed to enter it.

4.1 A generalized model sensitive to preemption

The central idea is that the model proposed by Cheng and Novick (2005) can be extended so that those occasions on which the target cause was preempted by an alternative cause are *attributed away* from the target cause. It is sufficient to focus

again on the situation where the potential cause factor C and the target effect factor E are both known to have occurred. The refined equation computing $P(c \rightarrow e|c, e)$ I would like to propose is:

$$P(c \rightarrow e|c, e) = \frac{w_C - w_C \cdot b_A \cdot w_A \cdot \alpha}{w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A} = \frac{w_C \cdot (1 - b_A \cdot w_A \cdot \alpha)}{P(e|c)}. \quad (4.1)$$

As can be seen, this equation extends Equation 3.2 by a new term that was added to the numerator, $w_C \cdot b_A \cdot w_A \cdot \alpha$. This term is intended to capture the probability with which the target cause was preempted by the alternative cause on an occasion. The product of w_C , b_A , and w_A identifies the sufficiency overlap of the potential causes. This product is relevant because, as we have seen earlier, the possibility of causal preemption can occur only on occasions on which the potential causes are simultaneously powerful enough to generate the effect (see Figure 3.3). The new parameter α represents an allocation parameter, $0 \leq \alpha \leq 1$, which partitions the sufficiency overlap of the potential causes such that the proportion of the sufficiency overlap is determined that needs to be “taken away” from the target cause. As the product $w_C \cdot b_A \cdot w_A \cdot \alpha$ is supposed to capture the probability of preemption of the target cause by the alternative causes, it has to be subtracted from w_C in the numerator. For example, the extreme case in which a reasoner is certain that an alternative cause preempted the target cause in its efficacy on an occasion on which target and alternative cause were simultaneously sufficient can be modeled by setting the α parameter to 1. By setting $alpha = 1$, the whole sufficiency overlap of the potential causes is attributed away from the target cause. The other extreme case in which a reasoner is certain that alternative causes did not preempt the target cause on an occasion of simultaneous sufficiency can be modeled by setting α to 0. In this case, Equation 4.1 reduces to the standard model (Equation 3.2) proposed by Cheng and Novick (2005).

4.2 Temporal information determines the Alpha parameter

The central idea of the new model is that the value of the α parameter is determined by knowledge about the temporal relation between the potential causes of the effect. For simplification, let us focus first on the type of situation that is given in the philosophical standard examples of preemption, which concerns scenarios in which the set of potential causes of the target effect contains only two elements, the target

cause factor C and one alternative cause factor A , and in which both potential cause factors are known to have been instantiated (in this case, the base rate parameter of the alternative cause can be set to 1 in Equation 4.1 and $P[c \rightarrow e|c, e]$ becomes $P[c \rightarrow e|c, a, e]$), and their causal powers w_C and w_A have already been learned.

4.2.1 Onset difference

The rock-throwing scenario introduced earlier suggests that one type of temporal information that should be relevant for the assessment of the probability of causal preemption is information about the onset time difference of the potential causes. Everything else being equal, if a occurs earlier than c on an occasion on which both happen to be probabilistically sufficient for the effect, then a can be assumed to preempt c in its efficacy. Conversely, c can be assumed to preempt a when c occurs earlier than a on an occasion of simultaneous sufficiency. Formally, the instantiation times of two potential causes c and a can be denoted t_c and t_a , and the difference between these onset times can be denoted $\Delta_t = t_a - t_c$. A schematic illustration using time indexed neuron diagrams is depicted in Figure 4.1 A). The onset times of the respective events is depicted on the x-axes in the Figure. It can be said that if two potential causes c and a are sufficiently powerful on an occasion, c preempts a in causing e when Δ_t is positive. Conversely, a preempts c in causing e when Δ_t is negative. A generic-level representation Δ_T can be obtained through averaging over multiple observations in which Δ_t is computed.

4.2.2 Causal latency

A second type of temporal information that is relevant for the assessment of the α parameter is information about the cause factors' causal latencies. By causal latency I mean the time that it takes a cause factor to unfold its causal power, which formally can be denoted $t_{C \rightarrow E}$ (cf. Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018). For example, taking an aspirin cannot be expected to bring relief right away. It first needs to be dissolved by gastric acid in the stomach, then must enter into the bloodstream, and finally must reach tissue where it can inhibit the production of prostaglandins. In our rock-throwing example, the causal delay with which Suzy's throwing her rock led to the bottle's shattering depended on the speed of the rock and the distance it had to travel until it reached its target. As another example, if we enter an elevator and press a button, we will notice that the doors do not shut immediately. Modern elevator systems possess a particular pre-programmed delay with which button presses become effective. Even pressing the on-button on the

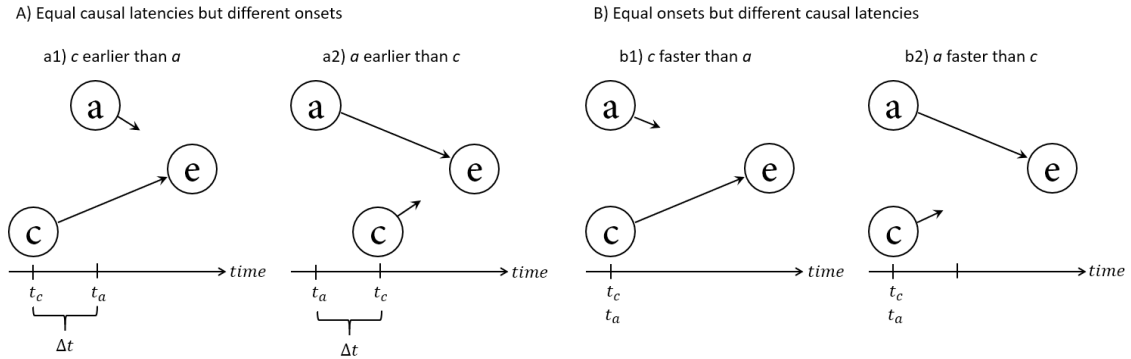


Figure 4.1: Illustration of the temporal components relevant for the assessment of the preemptive relation between two potential causes of an effect. A) illustrates the relevance of the instantiation times of the potential causes. B) illustrates the relevance of the causal latencies of the potential causes.

remote control does not turn on the TV instantaneously, the light signal first has to gap the distance between the remote control and the receiver.

An important characteristic of causal latencies is that they are, like causal powers, unobservable properties of causes. The causal latency of a cause factor C must be inferred from multiple observations of the onset differences between singular instantiations of C and E . Only in an environment that is shielded from potential alternative causes do the observed onset differences between C and E represent a direct measure of the causal latency of C , just like ΔP provides a direct measure of the causal power of a target cause factor only when the learning context is shielded from the influence of alternative factors (i.e., when the base rate of the effect is zero). The remote control and the pain killer scenario are good illustrations for this. Whereas TVs rarely turn on if we do not press the button on the remote control, headaches typically subside after a while even without the help of aspirin. Both the causal power as well as the causal latency of the painkiller are harder to learn than the causal power and the causal latency of the remote control.

Neuron diagrams illustrating the relevance of causal latency for preemption are depicted in Fig. 4.1 B. The figure illustrates that, everything else being equal, a probabilistically sufficient potential cause c of the effect e preempts a simultaneously sufficient potential alternative cause a on an occasion if c 's causal latency is smaller than a 's. Conversely, a preempts c in causing e if a 's causal latency is smaller than c 's on such an occasion. Causal latency here refers to the hypothetical time it would have taken a cause to produce the effect if it had acted in isolation.

Of course, the “everything else being equal” criterion will only rarely, maybe never, hold in real life situations. On most occasions different onset times will be combined with different causal latencies, which leads to the possibility that causes

with high causal latency can still preempt alternative causes if they have an onset advantage on their side. Likewise, causes that occur after their competitors can still manage to preempt them if they have a small causal latency. Onset difference and causal latency are assumed to be compensatory temporal factors.

A number of studies that have investigated the role of time in causal reasoning (e.g., Bramley et al., 2018; Buehner & May, 2003; Griffiths & Tenenbaum, 2009; Hagmayer & Waldmann, 2002; Lagnado & Sloman, 2004, 2006) have shown that reasoners are sensitive to causal latency (see Buehner, 2017, for a recent review). In an early study, Shanks, Pearson, and Dickinson (1989), for example, showed that subjects who experienced a noticeable temporal gap between a potentially causal action and the occurrence of the potential effect had lower impressions of a causal connection between action and outcome than subjects who experienced that outcome succeeded the action in a highly contiguous manner, in line with the idea that short causal delays are advantageous if everything else can be assumed to be constant. Later studies revealed that the contiguity effect indeed interacts with reasoners' background assumptions about the causal mechanism relating cause and effect factor (e.g., Buehner & McGregor, 2006; Hagmayer & Waldmann, 2002). In studies in which subjects were provided with causal mechanism information that made a delay plausible, the preference for short delays either diminished or sometimes even reversed. For example, in one study Buehner and McGregor (2006) confronted subjects with an unfamiliar apparatus in which marbles running down a path way could activate different switches connected to a light bulb. During an inspection phase, the steepness of the pathways within the machine was varied and thus either suggested a relatively small or a relatively large delay between marble insertion and activation of light. During the test phase, the mechanism part of the device was covered and subjects observed multiple instances of marble insertion and light activation. The observed delays were either compatible or incompatible with previously learned causal mechanism. The results showed that subjects who observed highly contiguous successions of events only gave high causality ratings when the mechanism in fact suggested short delays, while causality ratings were low when subjects expected the delay to be longer. Hagmayer and Waldmann (2002) showed that background assumptions about causal delays constrain causal learning based on purely statistical information.

The results of a study conducted by Lagnado and Speekenbrink (2010) provides evidence that the on average adverse effect of relatively high causal latencies on causal judgments might indeed be explained by subjects' assumptions about the preemptive relation between target cause and potential alternative causes in

the observed singular instances during learning. In their experiment, Lagnado and Speekenbrink (2010) varied the causal latency of the target cause factor and the probability with which alternative causes of the effect occurred during the cause-effect interval. This experimental design allowed to compare causal judgments between conditions with equal causal latencies but varying probabilities with which potential alternative causes occurred *before* the instantiation of the effect. Causal judgments tended to be lower when the probability of alternative causes occurring before the onset of the effect was high but not when this probability was low. An explanation for this difference is that the occurrence of alternative causes before the occurrence of the effect opens up the possibility that the target cause was preempted, while on occasions on which the alternative causes tended to occur after the effect preemption of the target cause can be safely ruled out. Moreover, Lagnado and Speekenbrink (2010) showed that variation in causal latency alone did not affect causal judgments.

4.2.3 Modeling causal latencies

The examples discussed above about the elevator or about the light traveling from the remote control to the receiver show that it might sometimes be appropriate to represent the causal latency of a cause factor as a stable, fixed value. In most contexts, however, causal latencies will be much less reliable and exhibit variation. The Aspirin case provides an example for such a context. A dose of medication might be expected to take effect after about twenty minutes, but we would probably not be too surprised if the effect occurred a bit earlier or a bit later. In most contexts it thus seems appropriate to represent the causal latency of a factor not as fixed value but as a random variable following a particular distribution. Variation in causal latency might be causally explained by the operation of unobserved hasteners or delayers (cf. Lagnado & Speekenbrink, 2010). For example, the time it takes an Aspirin to bring relief on a singular occasion is influenced by the state of various factors, such as the momentary concentration of gastric acid in the person’s stomach, her momentary blood pressure, and so on.

A standard way of modeling latencies that exhibit variation is to use gamma distributions. Gamma distributions are, for example, used in queueing theory to model waiting times (Shortle, Thompson, Gross, & Harris, 2018). Bramley et al. (2018), who investigated the role of time in general-causal structure induction, used gamma distributions in their study to model the causal latency of cause factors. The gamma distribution belongs to the family of continuous probability distributions and represents a generalization of the exponential distribution. It is characterized by two parameters: shape, $\kappa > 0$, and scale, $\theta > 0$. If gamma distributions are used to model

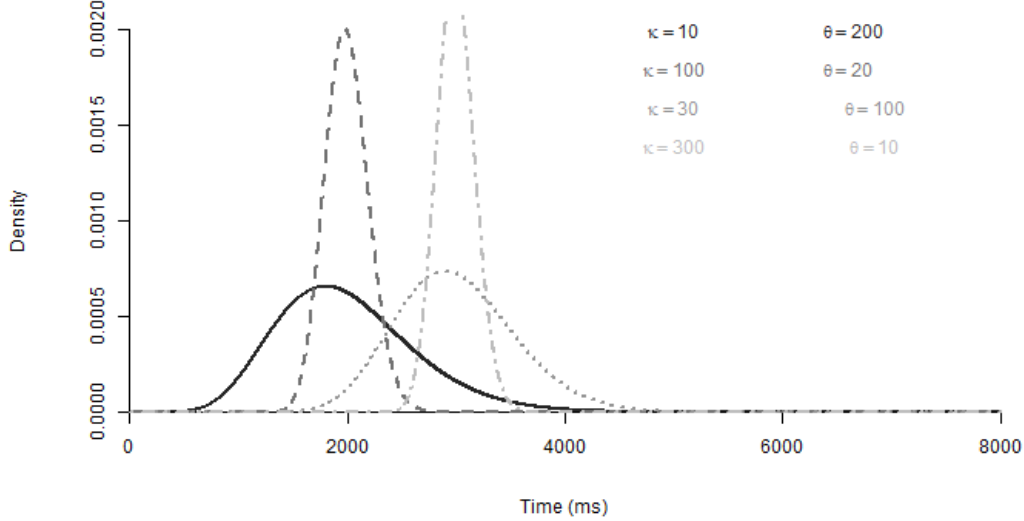


Figure 4.2: Example of gamma distributions with different shape (κ) and scale (θ) parameters. The expected value of a gamma distribution is given by $E[X] = \kappa \cdot \theta$. The expected values of the example distributions are illustrated by the vertical lines.

causal latencies, the expected value of the causal latency of a cause factor, $t_{X \rightarrow Y}$, is $E[t_{X \rightarrow Y}] = \kappa \cdot \theta$. The variance of the causal latency is $Var[t_{X \rightarrow Y}] = \kappa \cdot \theta^2$. Different gamma distributions that are supposed to depict the causal latencies of different cause factors are shown in Fig. 4.2 (temporal units represent ms in this example). The higher the expected value of a distribution, the slower a cause is expected to generate its effect. The larger the variance of a distribution, the less predictable the time of the occurrence of the effect becomes. Extremely reliable causal latencies, i.e., distributions that are highly centered around $E[t_{X \rightarrow Y}]$, can be modeled by holding $E[t_{X \rightarrow Y}]$ fixed and letting $\kappa \rightarrow \infty$. In Figure 4.2, the left two causal latency distributions have an expected value of $E[t_{X \rightarrow Y}] = 2000ms$ but different variability and the right two distributions have an expected value of $E[t_{X \rightarrow Y}] = 3000ms$ but different variability.

4.2.4 A formalization of the Alpha parameter

The formalization of the different temporal components that influence the probability of preemption allow it to quantify α for different types of situations. I will focus again on the type of situation that we have already considered above, in which it is known that the two potential causes factors C and A and the effect factor E have co-occurred (c, a, e) on an occasion. Moreover, we consider contexts in which the causal latencies of the potential cause factors $(t_{C \rightarrow E}, t_{A \rightarrow E})$ and the onset difference

on the particular occasion (Δ_t) have been learned. In this case the probability that the target cause c was preempted by the alternative cause a if both happen to be simultaneously sufficient for the effect corresponds to:

$$\alpha = P(t_{a \rightarrow e} + \Delta_t < t_{c \rightarrow e} | e, c, a). \quad (4.2)$$

In this type of situation α is the probability that the sum of the causal latency of the competing cause factor A and the time lag between c 's and a 's onset times is smaller than the causal latency of C , given e, c, a .

4.2.5 Approximating alpha with a Monte Carlo algorithm

When the causal latencies can be represented as fixed values, like in the elevator example, α can be determined by comparing the simple sum of $t_{a \rightarrow e}$ and Δ_t with $t_{c \rightarrow e}$. In this case, α will either be 0 or 1. In the more complicated case in which the causal latencies exhibit variability and are modeled with gamma distributions, α can take on any value between 0 and 1. In this case, the value of α can be estimated with the following Monte Carlo (MC) algorithm:

1. Sample N pairs of causal latencies ($t_{c \rightarrow e}, t_{a \rightarrow e}$) from the latency distributions ($t_{C \rightarrow E}, t_{A \rightarrow E}$) of C and A , respectively.
2. Calculate $t_{a' \rightarrow e} = t_{a \rightarrow e} + \Delta_t$ for all sampled $t_{a \rightarrow e}$ -values.
3. Count all pairs for which $t_{a' \rightarrow e} < t_{c \rightarrow e}$.
4. Divide this count by N .

For an illustration, let us assume that the causal latency of the target cause C , $t_{C \rightarrow E}$, followed the gamma distribution with the parameters $\kappa = 10$ and $\theta = 200$ in Fig. 4.2, and that the causal latency of the alternative cause A , $t_{A \rightarrow E}$, corresponded to the distribution with the parameters $\kappa = 30$ and $\theta = 100$. The expected value of the target cause's causal latency is $E = \kappa \cdot \theta = 2000 \text{ ms}$, while causal latency of the alternative cause has an expected value of $E = \kappa \cdot \theta = 3000 \text{ ms}$. Hence, the target cause tends to unfold its causal capacity quicker than its competitor. However, it can be seen that there is also an overlap between the two causal latencies. This overlap implies that, among all occasions on which both causes have identical onset times, there will be several cases in which the alternative cause outperforms the target cause. The α parameter captures the probability with which this happens. For occasions on which the two causes occur simultaneously (i.e., $\Delta_t = 0$) the MC algorithm presented above yields a value of $\alpha = 0.12$ for this pair of distributions.

That is, the probability that a acts quicker than c on a singular occasion on which the causes have identical onsets is twelve percent. Now consider the case in which the causal latencies are much more reliable and follow the two strongly peaked distributions. In this case, the probability that the target cause is preempted by the alternative cause reduces to $\alpha = 0.0001$ and becomes negligibly small. Finally, consider the case in which the target cause and the alternative cause not only occurred at the same time but in which their causal latencies followed the same gamma distribution. In this case where the two latency distributions fully overlap, the algorithm introduced above yields a value of α of 0.50, because in the set of randomly sampled pairs of causal latencies the causal latency of the alternative cause would be smaller in half of the pairs.

4.3 What if alternative causes are unobserved

We have thus far discussed the type situation in which there exist only two potential cause factors of the effect, C and A , that both are known to have been instantiated in the target situation. In this type of situation, temporal information about the onsets of the potential causes and about their causal latency modeled as gamma distributions should be used to determine α . A question is whether α can also be determined in situations in which there exist multiple alternative causes of the effect that have remained unobserved. To estimate α in this case, what we would need to represent is the temporal distribution of the base rate of the effect. One idea is that the temporal distribution of the base rate of the effect can be modeled with exponential distributions, which represents a special case of the gamma distribution in which the shape parameter κ is set to 1. Modeling the temporal distribution of the effect's base rate with exponential distributions seems appropriate because the exponential distribution has been described as "memoryless" (cf. Bramley et al., 2018), which means that the expected value with which the target outcome occurs over time remains constant. It does neither depend on previous occurrences of the target outcome nor on the onset times of the generative process. All other types of gamma distributions, by contrast, introduce a temporal dependence of the outcome on the modeled onset time of the generative process. If the temporal distribution of the effect's base rate is modeled with an exponential function and the causal latency of the target cause follows a known gamma distribution, Δ_t can be set to 0 and the same sampling algorithm as before can be used to determine the value of α . A graphical illustration of how different values of α can result in this type of context is shown in Figure 4.3, where two different gamma distributions ($\kappa = 10$, $\theta = 100$

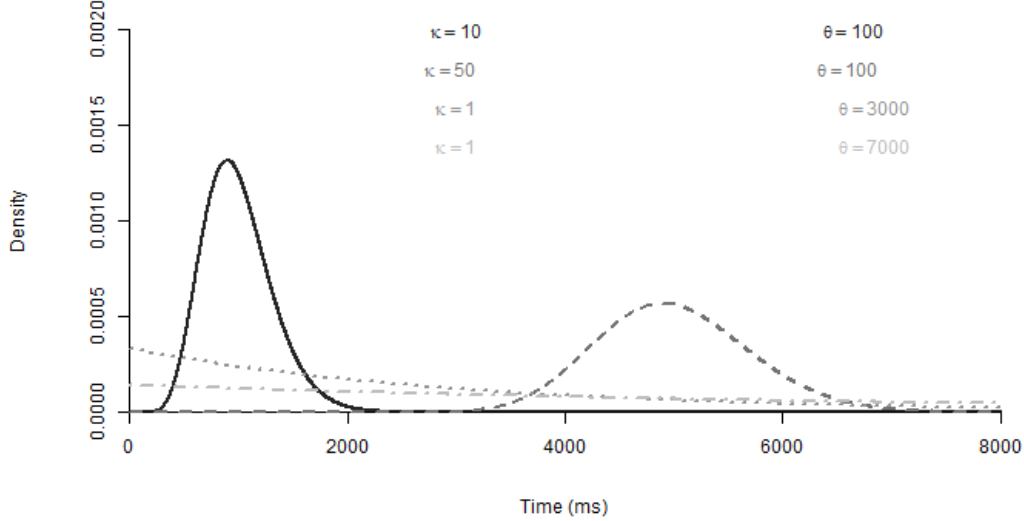


Figure 4.3: Examples of gamma and exponential distributions. Exponential distributions can be used to model the temporal distribution of the base rate of the effect when alternative causes of the effect are unobserved.

and $\kappa = 50$, $\theta = 100$) and two different exponential distributions ($\kappa = 1$, $\theta = 3000$ and $\kappa = 1$, $\theta = 7000$) are depicted. For example, in a context in which the temporal distribution of the base rate of the effect follows the exponential distribution with the parameters $\kappa = 1$ and $\theta = 3000$, an effect can be expected to occur after three seconds whenever an unobserved alternative cause is sufficiently powerful to generate the effect on an occasion. Imagine that the target cause’s causal latency followed the left gamma distribution in the Figure ($\kappa = 10$, $\theta = 100$). For this pair of distributions, α would be 0.28, indicating that the target cause tends to be preempted in its efficacy of producing the effect on about thirty percent of all occasions on which the target and an alternative cause are simultaneously powerful enough. If the target cause acted with a much slower latency of $\kappa = 50$ and $\theta = 100$, this probability would increase to $\alpha = 0.81$. Now consider the case where the base rate distribution followed the exponential distribution with $\kappa = 1$ and $\theta = 7000$. In this case, if the target cause’s causal latency followed the first gamma distribution, α would be 0.13, whereas it would be 0.51 its causal latency followed the second gamma distribution. This example illustrates in a formal way why small causal latencies of the target cause are always “preferable” in contexts in which alternative causes remain unobserved.

4.4 What if temporal information is not available

Another question that arises in the context of the present analysis is what should be done in the absence of any temporal information. For example, in our scenario about the fictitious medicine that can cause headaches as a side effect, all information that was available was statistical information that allowed to compute the causal powers of the potential causes. We have seen in this chapter that α should take on a value of 1 to model situations in which a reasoner is certain that the target cause was preempted by an alternative cause. One idea for situations in which temporal information is absent and in which a reasoner is uncertain about the preemptive relation of the potential causes is to set α to 1 here, too. In a context in which temporal information is absent, the output of Equation 4.1 obtained with α being set to 1 can be interpreted as an estimation of the lower boundary of the probability with which c caused e on an occasion. $P(c \rightarrow e|c, e)$ obtained by setting $\alpha = 1$ can thus be regarded as a conservative estimate of the probability with which c caused e . Another possibility would be to compute $P(c \rightarrow e|c, e)$ under $\alpha = 1$ and $\alpha = 0$ and to determine the range between the two values. Under a Bayesian interpretation, the estimation obtained under $\alpha = 1$ would represent the minimal degree of belief that c caused e and the estimation obtained under $\alpha = 0$ would represent the maximal degree of belief in c having caused e on the target occasion. Equation 4.1 shows that the range between the two values is determined solely by the size of the causal power overlap of the potential causes.

4.5 Summary

The present chapter introduced a generalization of Cheng and Novick's (2005) Power Framework of Causal Attribution that incorporates the possibility of causal preemption of the target cause by an alternative cause. The refined model that was proposed contains a new term that represents the probability with which the target cause was preempted. It was argued that the assessment of the probability of preemption requires the combination of two types of information: information about causal power and temporal information. Information about causal power is relevant because the problem of preemption cannot occur unless multiple causes are simultaneously powerful enough to produce the effect. On occasions of simultaneous sufficiency, temporal information is necessary to determine whether the target cause was preempted or not. Temporal information is supposed to be captured by a newly introduced parameter α . Two types of temporal information that are considered to be relevant for the determination of α were then introduced; information about the

potential causes' onset times and about their causal latencies. Causal latency is the time it takes a cause to produce its effect, which must be inferred from multiple observations that track the delay between cause and effect onset. A formalization of the α parameter was then provided that allows it to test quantitative predictions of the new model. In Chapter 6 I will summarize the results of a set of studies (Stephan et al., submitted, see Appendix B) in which crucial predictions of the new model were empirically tested. Finally, for contexts in which temporal information is not available, it was suggested that α should take on a default value of 1. The reason is that the predictions obtained by the new model in this case can be interpreted as a conservative estimate of the probability with which c caused e on the target occasion.

Chapter 5

Incorporating General Causal Structure and Parameter Uncertainty

“We sail within a vast sphere, ever drifting in uncertainty, driven from end to end.”

- Blaise Pascal

The incapability of the standard Power Framework of Causal Attribution to handle causal preemption was characterized in the previous chapter as a *conceptual* problem of the framework, which, as we have seen, is due to the fact that the framework attributes causality to the target cause whenever the target cause can be assumed to be sufficiently powerful for the generation of the effect on an occasion. We have seen in the last chapter that this conceptual shortcoming can be remedied by incorporating temporal information. Another problem of the Power Framework of Causal Attribution is that its predictions are based solely on *point estimations* of the causal power parameters obtained directly from the observed data (see Griffiths & Tenenbaum, 2005, 2009; Holyoak, Lee, & Lu, 2010; Meder & Mayrhofer, 2017; Meder et al., 2014). The framework thus fails to take into account that sample-based inferences are always associated with varying degrees of uncertainty (we have seen in Chapter 2 that, as a consequence, causal power cannot, for instance, explain different intuitions about zero-contingencies in which the base rate of the effect varies). The focus of this chapter will be to show how inferential uncertainty can be incorporated by the model. The result is an extension of the model that I developed based on the Structure Induction Framework proposed by Meder et al. (2014). The resulting model is called the Structure Induction Model of Singular Causation Judgments.

Table 5.1: Different sets of observations for which the Power Framework of Causal Attribution predicts maximum values of $P(c \rightarrow e|c, e)$ because the effect is sometimes observed in the target cause’s presence but never in its absence.

	Observations				Point estimations		
	$N(c, e)$	$N(c, \neg e)$	$N(\neg c, e)$	$N(\neg c, \neg e)$	w_C	$b_A \cdot w_A$	$P(c \rightarrow e c, e)$
Data Set 1	2	6	0	8	0.25	0	1
Data Set 2	4	4	0	8	0.50	0	1
Data Set 3	6	2	0	8	0.75	0	1
Data Set 4	24	24	0	48	0.50	0	1

To illustrate the relevance of “inferential uncertainty” for judgments of singular causation, let us consider the different contingency data sets that are listed in Table 5.1. We can again assume that these were the results of different fictitious studies investigating the causal capacity of a medicine to lead to undesirable side effects. In all data sets the effect was never observed in the absence of the cause ($n[-c, e] = 0$), whereas the probability with which it occurred in the cause’s presence, $P(e|c)$, increases from 0.25 in Data Set 1 to 0.50 in Data Set 2, to 0.75 in Data Set 3. Data Sets 1 to 3 contain 16 observations. In Data Set 4, 48 cases were observed and $P(e|c) = 0.50$, like in Data Set 3. The point estimations of the relevant causal parameters are depicted in the right part of Table 5.1. Imagine we had observed the contingency contained in Data Set 1 and now encountered a singular case in which C and E have been co-instantiated (c, e). How confident would we be under these conditions that c was the singular cause of the observed e ? The right column of Table 5.1 shows the predictions of Cheng and Novick’s (2005) Power Model of Causal Attribution. We see that the model predicts that we should be certain that c and e were causally connected in this situation, as $P(c \rightarrow e|c, e) = 1$. Importantly, the generalized model that was introduced in the previous chapter makes identical predictions in this case, as the new term in the numerator of Equation 4.1 is zero. We have earlier seen that the model makes this prediction here because it assumes that alternative cause factors of the effect do not exist in the present context ($b_A \cdot w_A = 0$) and that the target cause factor C is necessary for the generation of the effect. The problem that has thus far not been addressed is how confident we can be based on the given sample of observations that C really has the causal power suggested by the point estimations obtained from applying Equation 2.6 to the data, and that alternative causes are indeed not at play (i.e., that $b_A \cdot w_A = 0$). After all, the sample of Data Set 1 contains only 16 observations and the observed effect is rather small, which suggests that we cannot be too confident that the point estimate of 0.25 for C ’s causal power is reliable. The observed results do not seem to be too

unlikely to have occurred from mere *sampling variation*, which would imply that the true causal power of the target cause is zero and that the observed effects were actually produced by the alternative causes. Intuitively, this uncertainty regarding the reliability of the sample information should lead to a more cautious singular causation judgments. Now imagine we had instead observed the higher contingency of 0.75 in Data Set 3. The model makes the same prediction in this case but, due to the higher empirical effect, the data seem to provide better evidence that the target cause possesses the causal power to generate the effect and that alternative causes can be ruled out in the given context. Now consider Data Sets 2 and 4. In both cases the contingency is 0.50 but the sample of Data Set 4 contains six times as many observations and should thus be regarded as more reliable.

What the previous example cases demonstrate is that our degree of belief in a singular causal connection between an observed co-instantiation of target cause factor C and target effect factor E seems to depend on how strongly the assumptions we make about the *general* causal relation between C , A , and E are supported by the observed data. Importantly, it has been argued (cf. Griffiths & Tenenbaum, 2005; Meder et al., 2014) that these assumptions may concern both the causal parameters that enter into the equations as well as the general causal structure assumed to underlie the observed data.

5.1 Incorporating causal parameter uncertainty

To account for the problem of inferential uncertainty, Holyoak et al. (2010) have proposed a Bayesian variant of Cheng and Novick’s (2005) Power Model of Causal Attribution (Equation 3.2). Unlike the standard model that merely relies on point estimations of the causal parameters that are derived directly from the data, the Bayesian extension of the model proposed by Holyoak et al. (2010) uses probability distributions over the causal parameters of the general common-effect causal structure assumed by Cheng’s (1997) Causal Power Theory. The model incorporates inferential uncertainty about the causal parameters by using uninformative, flat prior distributions over the causal structure’s parameters that are updated in light of the observed contingency data.

5.2 Incorporating general-causal structure uncertainty

Recent work by Meder et al. (2014) suggests, however, that the incorporation of uncertainty about the causal parameters of a fixed general causal structure might

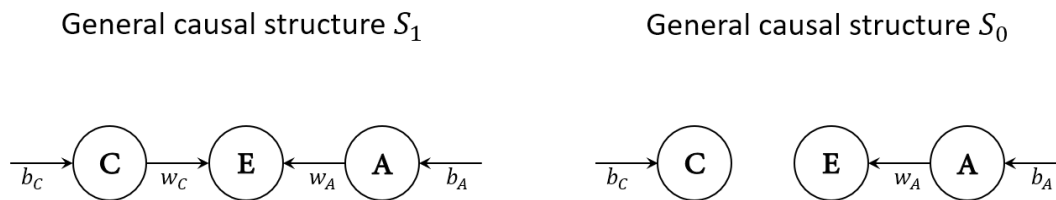


Figure 5.1: The two causal structures considered by the Structure Induction Framework as mutually exclusive explanations for observed patterns of covariation. C is a general cause of E under S_1 , whereas all co-occurrences of C and E are coincidental under S_0 .

not be sufficient to account for people’s causal inferences (though see Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). Relying on previous work by Griffiths and Tenenbaum (2005), Meder et al. (2014) developed and tested a new computational model, the Structure Induction Model, which in addition to parameter uncertainty also incorporates uncertainty about the general *causal structure*. The two general causal structures considered by the Structure Induction Model as alternative causal hypotheses explaining the observed data are shown in Figure 5.1 (see also Griffiths & Tenenbaum, 2005). Structure S_1 is the common-effect structure assumed by the standard Causal Power Model and by the Bayesian attribution model of Holyoak et al. (2010). S_0 represents a causal structure in which the causal arrow connecting C and E is missing. S_0 captures the causal hypothesis according to which all observed co-occurrences of C and E are coincidental and all observed effects are actually caused by alternative causes A . The framework uses Bayesian inference over causal models (see Griffiths, Kemp, & Tenenbaum, 2008, for an overview) to assess which general causal structure is more likely to account for the observed data.

Meder et al. (2014) did not investigate general causal structure and parameter uncertainty in singular causation judgments, however, but in diagnostic judgments. Diagnostic judgments are formally captured by $P(c|e)$. Unlike singular causation judgments that express how likely it is that an observed effect was *actually caused* by a target cause factor C (denoted by the causal arrow between c and e in $P[c \rightarrow e|e]$ or $P[c \rightarrow |c, e]$), diagnostic judgments refer to the probability with which a potential cause factor C was *co-instantiated* with an observed effect e on an occasion. According to the Causal Power Theory, $P(c|e)$ also includes cases in which e was actually caused by an alternative cause. Meder et al. (2014) demonstrated that reasoners’ diagnostic judgments are influenced not only by the empirical probability of $P(c|e)$ computed directly from the data, but also by the degree to which the observed data support the hypothesis that C and E are generally causally linked, formally captured by the probability that structure S_1 is underlying the data (an

elaborate overview is given by Meder & Mayrhofer, 2017). The idea incorporated by the framework is that an observed instantiation e of the effect factor provides evidence for the presence of C only if C can be assumed to be a cause of E . A model that incorporates only parameter uncertainty, by contrast, makes the default assumption that C can generally cause E .

Even though the Structure Induction Framework has originally been developed to account for general structure and parameter uncertainty in diagnostic judgments, it can also be used to incorporate general causal uncertainty into predictions of singular causation judgments. The motivation to incorporate general causal structure and parameter uncertainty into estimations of singular causation is the assumption that the degree to which a reasoner believes that c caused e on a singular occasion should be constrained by the degree to which she believes that C generally possesses the causal capacity to produce E . In the following sections I describe how singular causation judgments modeled by $P(c \rightarrow e|c, e)$ can be derived within the Structure Induction Framework.

5.3 The Structure Induction Model of Singular Causation Judgments

To incorporate general causal structure and parameter uncertainty into an estimation of $P(c \rightarrow e|c, e)$, the Structure Induction Model carries out different computational steps. I will here focus on the core ideas behind these computational steps. The detailed analytic derivations can be found in Meder et al. (2014) or in Griffiths and Tenenbaum (2005).

5.3.1 Estimating the general causal structures' probabilities

One step of the model is to compute the probability with which the observed contingency data were generated by structure S_1 as opposed to structure S_0 . The probability of S_1 being the general causal structure underlying the data can be understood as the degree to which a reasoner believes that C can generally cause E . To estimate this probability, the model applies Bayes rule:

$$P(S_i|D) = \frac{P(S_i) \cdot P(D|S_i)}{P(D)} \text{ with } P(D) = \sum_{i \in \{0,1\}} P(D|S_i) \cdot P(S_i). \quad (5.1)$$

$P(S_i|D)$ denotes the posterior probability of general causal structure S_i . $P(S_i)$ denotes a structure's prior probability that is updated in light of the observed con-

tingency data D . To express initial uncertainty about the underlying causal structure, the model assigns an uninformative prior probability of 0.5 to each structure. $P(D|S_i)$ is the likelihood of the observed data under a particular structure. It thus indicates how much evidence the data provide for a particular structure. Formally, it is given by the integral over the likelihood function of the respective causal structure’s parameters (see below and cf. Meder et al., 2014). $P(D)$ represents the probability of the data under any of the two potential causal structures and serves as a normalizing constant.

Conceptually, given a fixed sample size of observations, the higher the observed contingency between the target cause factor C and the target effect factor E , the higher the probability of S_1 becomes, while the probability of S_0 as an explanation of the data is proportionally weakened. According to the likelihood function in Equation 5.1, the probability of observing a high statistical dependency between C and E is more likely if a causal arrow exists between C and E than if a causal arrow is missing. Hence, in Table 5.1, Data Set 3 provides better evidence for S_1 than Data Set 2 and Data Set 2 provides better evidence for S_1 than Data Set 1. In fact, for the weak contingency indicated by Data Set 1 in Table 5.1, the model estimates a posterior probability $P(S_1|D)$ of Structure S_1 of only around 0.5, indicating that the small positive effect observed in the data was not big enough to noticeably change the prior belief in S_1 . For Data Set 2, by contrast, the posterior probability of S_1 is already 0.85 and for Data Set 3 it is almost 1.0. An illustration is shown in Figure 5.2. Interestingly, if the low contingency in Data Set 1 had been even lower, the model would have favored S_0 over S_1 . S_1 would have been inferentially “punished” by the model for assuming an additional parameter (w_C) that S_0 is not assuming. This principle incorporated by the Structure Induction Framework is called *Bayesian Ockham’s razor* (see MacKay, 2003).

A second factor influencing the posterior probability of a causal structure is the sample size of observations. Given a fixed positive empirical contingency between C and E , the relative probability of S_1 increases with the number of observations that have been made. For example, in Data Sets 2 and 4 in Table 5.1 the contingency is 0.50 but the sample of Data Set 4 is six times higher than the sample of Data Set 2, yielding a posterior probability for S_1 of almost 1.0.

5.3.2 Estimating each potential structure’s set of parameters

A further computational step of the model is the estimation of each of a structure’s causal parameters. Importantly, unlike the standard Causal Power Theory (Cheng, 1997), the Structure Induction Model does not derive point estimations but com-

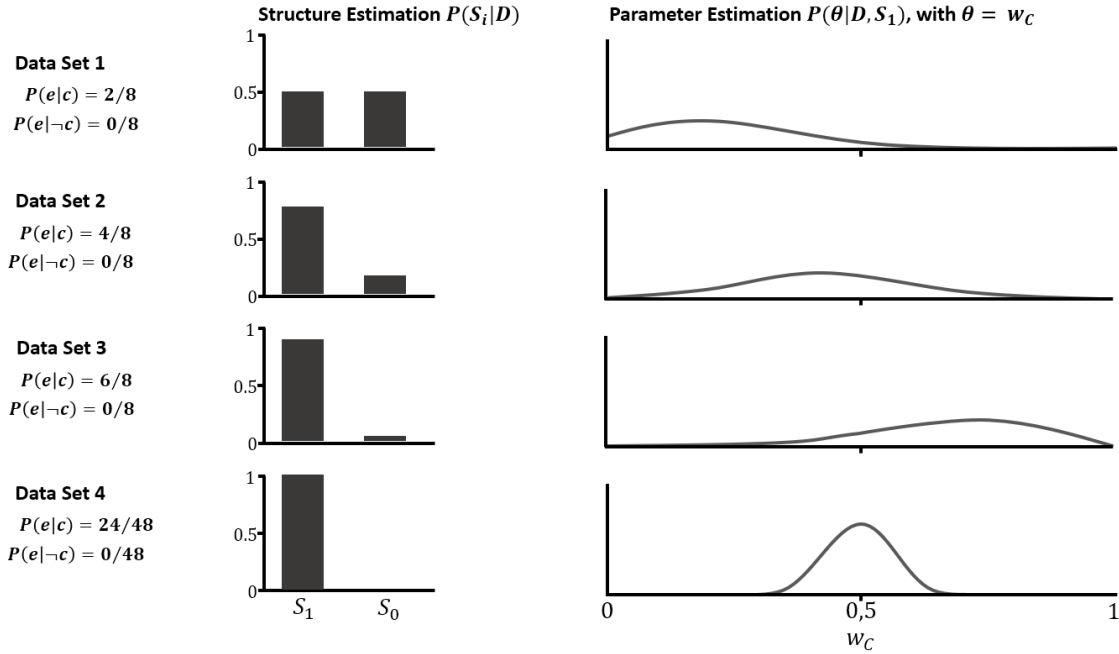


Figure 5.2: The two core inferential steps of the Structure Induction Model: Based on the observed contingency data D , the model estimates the posterior probability of S_1 and S_0 (left) as well as their parameters θ . The right part shows the posterior probability distribution over w_C of S_1 .

putes probability distributions over the structures' parameters. The distributions of the parameters' are estimated separately for each causal structure. As the model considers the default case in which alternative causes A are unobserved, the distribution of parameter b_A cannot be estimated independently of the distribution of causal power parameter w_A , which is why the parameters b_A and w_A are subsumed by the model under a single distribution w_A . Hence, under S_1 , distributions over b_C , w_C , and w_A are estimated. Under S_0 , by contrast, only the distributions over b_C and w_A are estimated, whereas the value of w_C is set to zero. To estimate the distribution over each of a structure's parameter, the model relies on Bayes rule:

$$P(\theta|D, S_i) = \frac{P(\theta) \cdot P(D|\theta, S_i)}{P(D|S_i)}. \quad (5.2)$$

$P(\theta|D, S_i)$ denotes the posterior probability density distribution over each of the structure's parameters. $P(\theta)$ is the prior probability density distribution over the parameters. To incorporate initial uncertainty about the parameter values, the model uses uninformative flat (Beta[1,1]) distributions over the different parameters. $P(D|\theta, S_i)$ denotes the likelihood of the data given a set of parameters and a particular causal structure. The likelihood function that delivers $P(D|\theta, S_1)$ is based on an assumed noisy-OR parametrization of S_1 (Pearl 1988, see also Meder

et al. 2014) and is given by

$$\begin{aligned}
 P(D|\theta; S_1) &= [(1 - b_C)(1 - w_A)]^{n(-c, -e)} \cdot [(1 - b_C)w_A]^{n(-c, e)} \\
 &\quad [b_C(1 - w_C)(1 - w_A)]^{n(c, -e)} \cdot [b_C(w_C + w_A - w_Cw_A)]^{n(c, e)}.
 \end{aligned} \tag{5.3}$$

As w_C is set to zero under S_0 , the likelihood function yielding $P(D|\theta, S_0)$ simplifies and is given by

$$\begin{aligned}
 P(D|\theta; S_0) &= [(1 - b_C)(1 - w_A)]^{n(-c, -e)} \cdot [(1 - b_C)w_A]^{n(-c, e)} \\
 &\quad [b_C(1 - w_A)]^{n(c, -e)} \cdot [b_Cw_A]^{n(c, e)}.
 \end{aligned} \tag{5.4}$$

For an illustration, let us consider how the power parameter w_C of C under S_1 is updated in light of the data. Conceptually, the degree to which the probability mass of w_C shifts towards a specific value in light of the observed data correlates positively with the observed statistical dependency (see Figure 5.2). For example, Figure 5.2 shows that from Data Set 1 to Data Set 3, the probability mass of w_C gradually shifts to the right, as the observed contingency increases from 0.25 in Data Set 1 to 0.50 in Data Set 2 to 0.75 in Data Set 3. Data Set 3 provides better evidence that the causal power of C is greater than zero because the mild peak in the posterior density distribution of w_C is farther away from a value of zero than the mild peak in the distribution of w_C obtained under Data Sets 1 and 2. However, it can also be seen that all three distributions are still relatively flat, which indicate a relatively high uncertainty about the exact value of w_C . The reason why all three distributions are rather flat is that the sample of observations was quite small. Generally, the posterior distribution of a parameter becomes increasingly centered the larger the sample of observations becomes. For example, Figure 5.2 shows that the distribution of w_C is much denser around 0.50 under Data Set 4 than it is under Data Set 2. This behavior of the model formally captures the intuition that an estimation is generally more reliable if the data on which it is based are many than if they are few.

5.3.3 Estimating the probability of singular causation under each parametrized structure

Once the model has estimated the posterior probabilities of the respective structures and their parameters, it computes the probability with which an observed singular co-occurrence of c and e was causal, $P(c \rightarrow e|c, e)$, given each of the two potential general causal structures. This is done by integrating over the parameters' values

weighted by their posterior probability, which yields $P(c \rightarrow e|c, e; \theta, S_i)$. Importantly, as C is assumed to generally lack the causal capacity to generate the effect under S_0 , $P(c \rightarrow e|c, e)$ is held fixed at 0 under S_0 . To estimate $P(c \rightarrow e|c, e; \theta, S_1)$ the generalized Equation (Eq. 4.1) yielding $P(c \rightarrow e|c, e)$ that was introduced in the previous chapter is applied. The estimate for $P(c \rightarrow e|c, e)$ obtained under S_1 incorporates uncertainty about the general structures' parameters. If Equation 3.2 was applied in this step, the resulting estimation for $P(c \rightarrow e|c, e)$ under S_1 would correspond to the predictions for singular causation of the refined model proposed by Holyoak et al. (2010).

5.3.4 Obtaining a single estimate for singular causation

The previous step yields two estimations of $P(c \rightarrow e|c, e)$, one for each potential structure, with a fixed value of zero for S_0 . However, the goal is to obtain a single final estimate that incorporates both parameter uncertainty and general causal structure uncertainty. This is achieved in a last computational step in which the model “integrates out” the two potential causal structures to deliver a “Bayesian average” of the target probability:

$$P(c \rightarrow e|c, e; D) = \sum_{i \in \{0,1\}} P(c \rightarrow e|c, e; D, S_i) \cdot P(S_i|D). \quad (5.5)$$

Conceptually, the final output represents a weighted average of the probability of $P(c \rightarrow e|c, e)$, where the weighting factor corresponds to the general causal structures' posterior probabilities. As $P(c \rightarrow e|c, e; D, S_0)$ is assigned a value of zero, $P(c \rightarrow e|c, e; D)$ essentially represents the product of $P(c \rightarrow e|c, e; D, S_1)$ and S_1 's posterior probability $P(S_1|D)$. The product of $P(c \rightarrow e|c, e; D, S_1)$ and $P(S_1|D)$ instantiates the assumption that the degree of belief in a singular causal connection between C and E cannot be higher than the degree of belief that C can be a cause of E .

For illustration let us consider the predictions that the Structure Induction Model yields for the two different contingency data sets that were introduced in the beginning of this chapter. For the first data set, the model yields a value for $P(c \rightarrow e|c, e; D)$ of about only 0.35. That is, it predicts a value that is far lower than the point estimation of 1 that the original model makes. In fact, the Structure Induction Model concludes here that it is more likely that an observed singular co-occurrence of c and e was coincidental than causal. To see why this is, consider again the last computational step given by Equation 5.5, which essentially delivers the product of $P(c \rightarrow e|c, e)$ under S_1 and S_1 's posterior probability. The value of

$P(c \rightarrow e|c, e)$ under S_1 , $P(c \rightarrow e|c, e; D; S_1)$, computed by the model lies around 0.7 for Data Set 1. The reason why already this value is below the point estimation of 1 delivered by the original model is that uncertainty about the S_1 's causal parameters is incorporated into this estimation (Equation 5.2). Importantly, the value for $P(c \rightarrow e|c, e)$ obtained under S_1 corresponds to the final prediction that the model proposed by Holyoak et al. (2010) would make, as this model considers only parameter uncertainty. Under the Structure Induction Model, by contrast, the value for $P(c \rightarrow e|c, e)$ obtained under S_1 gets reduced further by the multiplication with the posterior probability of S_1 (see Equation 5.5). We have seen earlier that the posterior probability for S_1 is only about fifty percent under Data Set 1. That is, the obtained value of $P(c \rightarrow e|c, e; D, S_i)$ gets reduced further by almost one half, yielding the final prediction of about 0.35. Now let us consider what happens for Data Set 2. Under Data Set 2, it is almost certain that S_1 is the underlying causal structure, as $P[S_1|D] \approx 1$. Consequently, the estimation of $P(c \rightarrow e|c, e; D, S_1)$ gets almost not lowered by S_1 's posterior probability. The model's final prediction thus ends up being closer to the point estimation delivered by Equation 4.1. Generally, the more inferential uncertainty concerning the general structure and the parameters decreases the more the model's predictions approximate the point estimates delivered by the original equations.

5.4 Summary

This chapter focused on the problem that the singular causation predictions made by the standard Power Model of Causal Attribution (and those made by the generalized model introduced in the previous chapter) rely only on point estimations obtained directly from the data. The predictions therefore fail to take inferential uncertainty into account. As a result, the model tends to overestimate the probability with which an observed potential cause c actually caused an observed effect e . The data sets introduced in the beginning of this chapter served as an illustration of this problem. We have then seen that inferential uncertainty might concern two things: the size of the causal parameters that enter into the equations and the general causal structure that is assumed to underlie the observed data. It was argued that singular causation judgments should be sensitive to both. Based on previous work by Meder et al. (2014), I proposed a Structure Induction Model of Singular Causation Judgments that incorporates general causal structure and parameter uncertainty into the equations of the Power Framework of Causal Attribution. The core assumption that this extended model captures is that the degree to which a

reasoner can believe that c caused e on a singular occasion is constrained by her degree of belief that C can generally cause E . In the next chapter, I will summarize a set of experiments (Stephan & Waldmann, 2018, see Appendix A) that aimed at demonstrating the validity of the extended model.

Chapter 6

Summary of the Empirical Findings

“It was an experiment, I suppose.”

- Sam Gayton, *The Snow Merchant*

In this chapter I will summarize the results of the experiments that were reported in the two articles on which this dissertation is based. The overall goal of these experiments was to evaluate the psychological validity of the new Power Model of Causal Attribution that I developed and presented in the last two chapters. The experiments of the first article (Stephan & Waldmann, 2018) aimed at establishing rather generally that the two problems that Cheng and Novick’s (2005) standard model neglects, the problem of causal preemption and the problem of general structure and parameter uncertainty, are reflected by people’s singular causation judgments. The studies reported in this article were previously included in the proceedings chapter by (Stephan & Waldmann, 2017, see Appendix E). The idea that structure and parameter uncertainty should be incorporated into the model was first proposed in the proceedings chapter by Stephan and Waldmann (2016, see Appendix D). The experiments reported in the first article focused on the type of context that has originally been considered by Cheng and Novick (2005), i.e., on situations in which only static contingency information is presented to participants. The studies of the second paper (Stephan et al., submitted) went beyond and focused on the interplay between causal power information and information about the different temporal factors that the new model considers to be relevant for the assessment singular causation, which were introduced in Chapter 4. Two of the experiments were previously reported in the proceedings paper by (Stephan, Mayrhofer, & Waldmann, 2018, see Appendix F).

In my summaries of the different studies I will focus on core experimental aspects such as the experimental goal, the procedure, and the obtained results. Study details, like the exact phrasing of the test questions or the results of the statistical analyses, can be found in Appendix B and C.

6.1 Experiments of Stephan and Waldmann (2018)

The paper by Stephan and Waldmann (2018) reports two different experiments. Experiments 1a and 1b aimed at testing whether reasoners generally take the possibility of causal preemption into account when making singular causation judgments, even if they only receive static contingency information like the results of the fictitious medical study that were used as an illustration in the previous chapters. We focused on situations with purely static information because it is the type of situation that Cheng and Novick (2005) considered when they proposed their model. Experiments 1a and 1b aimed at testing the generalized equation proposed in Chapter 4 that incorporates the possibility of preemption. To test the new equation, we tried to minimize any potential influence of structure and parameter uncertainty on subjects' singular causation judgments. The demonstration of an influence of general causal structure and parameter uncertainty on singular causation judgments was the goal of Experiment 2.

6.1.1 Experiments 1a and 1b

The paradigm that we used in Experiments 1a and 1b was a classical causal induction paradigm. As cover story we used the gene expression scenario previously reported in Griffiths and Tenenbaum (2005). Subjects were asked to take the perspective of biologists who wanted to test if a particular chemical substance can cause the expression of a particular gene in laboratory mice. They were informed that they will observe two random samples of mice, one sample of 24 mice that remained untreated, serving as a control group, and one sample of 24 mice that was treated with the chemical substance. Following Griffiths and Tenenbaum (2005), it was mentioned in the instructions that the control is important because some mice might express the gene for natural reasons. The results of the fictitious study were presented in a summary format (see Figure 1 in Appendix A) similar to the illustrations used to describe the results of the fictitious medical study in the previous chapters (e.g., Figure 2.1).

Three different contingency data sets were manipulated between subjects. In all three contingency data sets, all mice in the treatment group showed the effect, i.e., $P(e|c) = \frac{24}{24}$, whereas the base rate of the effect, $P(e|\neg c)$, observed in the control group varied from $P(e|\neg c) = \frac{0}{24}$ in the first data set to $P(e|\neg c) = \frac{8}{24}$ in the second data set to $P(e|\neg c) = \frac{12}{24}$ in the third data set. This set of contingencies was chosen because Cheng and Novick's (2005) Standard Power Model of Causal Attribution and the generalized model that takes the possibility of preemption into account

(Equation 4.1) make different singular causation predictions in this case. For all three data sets, the standard model yields a value of 1.0 for the probability that C caused E in a singular case in which C and E were co-instantiated. These predictions (see Figure 6.1 b) of the standard model are obtained because the estimated causal power (w_C) of the chemical is 1.0 in all three data sets. The generalized attribution model, by contrast, predicts a negative influence of the observed base rate of the effect. The reason is that unless the newly introduced α parameter takes on a value of zero, a proportion of the sufficiency overlap between the target and the alternative causes (of the size of α) is subtracted from the causal power of C (see Equation 4.1). In the contingency data sets that we tested, the sufficiency overlap between target and alternative causes corresponds to the observed base rate of the effect. We hypothesized that subjects would make the default assumption in this scenario that the alternative causes tend to preempt the target cause, for it appeared plausible that natural factors that also can cause gene expression had been instantiated before the biologists decided to conduct their study. In the new model, this assumption about the preemptive relation between C and alternative factors A is captured by high values of the new α parameter. We set α to 1.0 in the model, yielding predictions that decrease from $P(c \rightarrow e|c, e) = 1$ for the first condition to $P(c \rightarrow e|c, e) = 0.75$ for the second condition to $P(c \rightarrow e|c, e) = 0.5$ for the third condition (see Figure 6.1 b). As the goal of Experiment 1 was to demonstrate the influence of assumptions about causal preemption on singular causation judgments and not the potential influence of sampling uncertainty, we used a sample size of observations that was sufficiently large to infer with high confidence in all three conditions that the chemical substance is *generally causally effective*. An analysis of the three data sets with the Structure Induction Model revealed that with a sample of $N = 48$ mice, the posterior probability $P(S_1|D)$ of Structure S_1 was close to 1 for all data sets, even for the third data set in which the observed contingency was only 0.5 (see Figure 6.1 a).

In Experiment 1a, subjects were asked two causal test queries after they had inspected the presented contingency data. Subjects were first presented with a general causation query asking how strongly they believed that the chemical substance can cause the expression of the gene. This question was included as a control question, to rule out that uncertainty about the general causal structure can account for predicted differences in the singular causation ratings. Based on the computed posterior probabilities of S_1 , we assumed that subjects in all conditions would indicate equally high degrees of belief in a general causal connection between C and E .

The second test query was the singular causation query. This test question

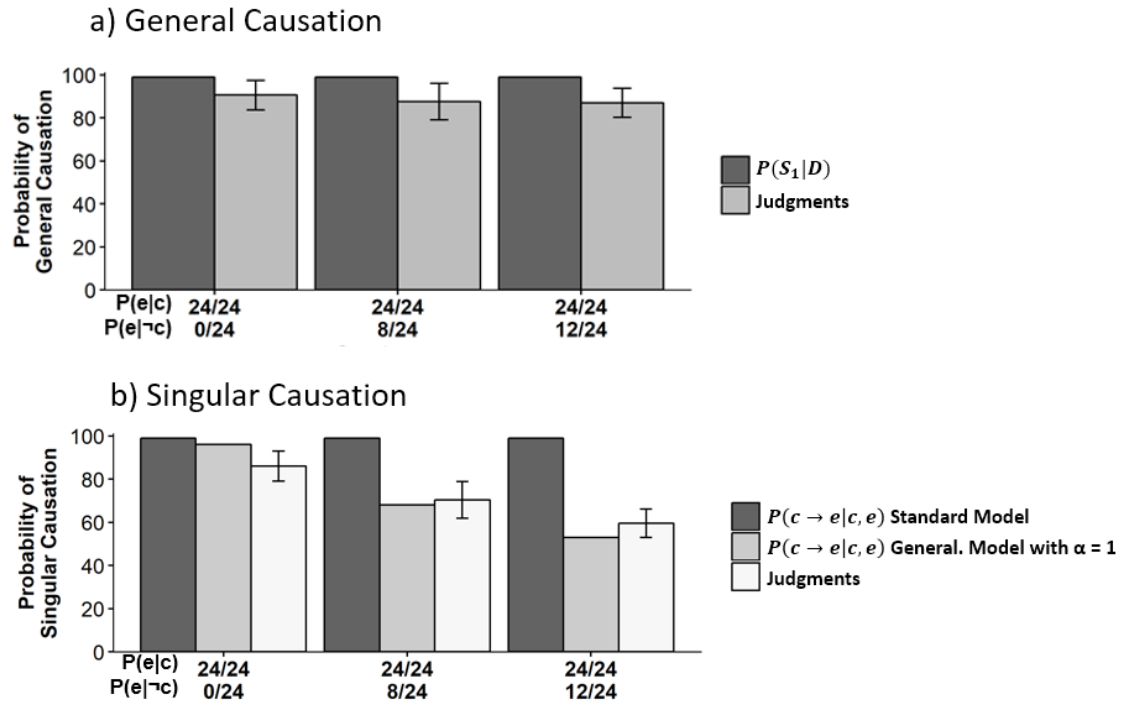


Figure 6.1: Predictions and results (means and 95% *CI*) of Experiment 1a (Stephan & Waldmann, 2018). Part a) shows general causation judgments for the three conditions together with the respective posterior probabilities of general causal structure S_1 as computed by the Structure Induction Model. Part b) shows singular causation judgments and the predictions made by the standard and the generalized Power Model of Causal Attribution.

referred to a randomly selected single mouse (called *Mouse #25*) from the treatment group (c) that was showing the effect (e). Subjects were asked to indicate how strongly they believed that it was the chemical substance that caused the expression of the gene in this particular mouse. Answers to this question were predicted to follow $P(c \rightarrow e|c, e)$, computed by the generalized model (Equation 4.1) with α set to 1.0. We thus expected to see decreasing singular causation ratings across the three conditions (see Figure 6.1).

The results (based on $N = 83$ subjects who were recruited from an online panel) are summarized in Figure 6.1. As for the control question about general causation, we found that subjects in all conditions were overall highly confident that the chemical generally possesses the capacity to bring about the effect (see Figure 6.1 a). The ratings for the general causation question did not differ significantly from each other across conditions. Subjects' singular causation ratings, by contrast, decreased from the first over the second to the third data set, as predicted by the generalized Power Model of Causal Attribution with α set to 1. Also, the ratings were close to the absolute values predicted by the model.

The goal of Experiment 1b was to obtain further evidence that the observed pattern of singular causation ratings was actually driven by subjects' assumptions about causal preemption. Like in Experiment 1a, subjects first answered the general causation question and then were presented with a single mouse from the treatment group that showed the effect. Yet, other than in Experiment 1a, subjects were not asked to say how strongly they believed that the chemical caused the effect in this mouse, but to indicate how likely they think it was that this mouse had already been expressing the gene before the chemical substance could have become effective. The question directly targeted the assumed probability of preemption of the chemical by background factors. Our prediction was that, if subjects assumed that the background causes tended to preempt the target cause, their ratings for this probability question should be close to the observed base rate of the effect (as our model predicts that preemption can only become an issue on occasions on which target and background causes are simultaneously sufficient). The results (based on $N = 74$ subjects recruited from an online panel) revealed that this was the case. Subjects' ratings were close to the observed base rate of the effect. As for the general causation query, we replicated the results of Experiment 1a. In all three conditions, subjects were highly confident that the chemical be generally causally effective.

Experiment 1 provided first evidence for the validity of the generalized model. When reasoners assess whether a potential observed cause c has generated an observed effect e on a singular occasion, they take into consideration that c might have been preempted in its efficacy by an alternative cause, even if they know that C is generally a very strong cause factor.

6.1.2 Experiment 2

The primary goal of Experiment 2 was to investigate whether we would find an influence of general causal structure and parameter uncertainty on reasoners' singular causation judgments, as predicted by the Structure Induction Model that was introduced in the last chapter. We also wanted to compare the new model's predictions to those made by different potential alternatives. To this end, we tested a larger set of contingency data in which different levels of $P(e|c)$ and $P(e|\neg c)$ were combined. The contingency data set¹ we tested was the one previously used by Buehner et al. (2003) to test Cheng's (1997) Causal Power Theory, and later by Griffiths and Tenenbaum (2005) to validate their Causal Support Model. The different contin-

¹The original set contained fifteen different contingencies but we left out the one in which the effect never occurs in C 's presence, as a singular causation query targeting $P(c \rightarrow e|c, e)$ cannot be asked if the effect is absent.

gencies are shown on the x-axis of the different graphs in Figure 6.2. It can be seen that the sample in each data set consisted of 16 observations and that $P(e|c)$ and $P(e|-c)$ were combined in a way that results in different levels of contingency (ΔP). In the first four data sets, the contingency is zero, followed by a contingency of 0.25 in Data Sets 5 to 8. In Data Sets 9 to 11 the contingency is 0.5. The contingency in Data Sets 12 and 13 is 0.75 and the contingency of the last data set is 1.0. General causal structure and parameter uncertainty was expected to decrease from the first to the last data set. The posterior probability of S_1 for each data set is shown in Figure 6.2 a). The singular causation predictions of our Structure Induction Model are depicted in Figure 6.2 c). Like in Experiment 1, we set the α parameter to 1. The right part of Figure 6.2 shows the singular causation predictions of different alternative models that we considered. The predictions of the Cheng and Novick's (2005) standard model (Equation 3.2) are shown in Graph (e). Graph (f) shows the predictions of the Bayesian version of Cheng and Novick's standard equation that was proposed by Holyoak et al. (2010), which incorporates only parameter uncertainty. Graph (g) shows the predictions of our Structure Induction Model with α set to zero in the generalized equation. As the generalized equation reduces to the standard model in this case, Graph (g) shows the predictions that Cheng and Novick's model would make if it incorporated structure and parameter uncertainty. Graph (h) shows the point estimates of $P(c \rightarrow e|c, e)$ obtained from the generalized attribution equation that was introduced in Chapter 4.

We used the same cover story and the same presentation format of the contingency information as in Experiment 1. We also again assessed both general and singular causation judgments. Other than in Experiment 1, however, the type of causal query (general vs. singular causation) was varied between-subjects, whereas the fourteen contingency data sets were varied within-subject. The cover story was therefore slightly modified. It additionally informed subjects that they will be presented the results of fourteen independent studies in which the effect of fourteen different chemicals on fourteen different genes were tested in fourteen different samples of mice.

The different contingencies were presented in random order and subjects could inspect each data set as long as they wanted. Subjects in the "general causation query" condition were asked to indicate how strongly they believed that the chemical they had just observed can cause the expression of the gene. They then proceeded to the next data set in which a different substance and a different gene were investigated. Subjects in the "singular causation query" condition were asked about a randomly chosen mouse from the treatment group (c) that showed the effect (e). The singu-

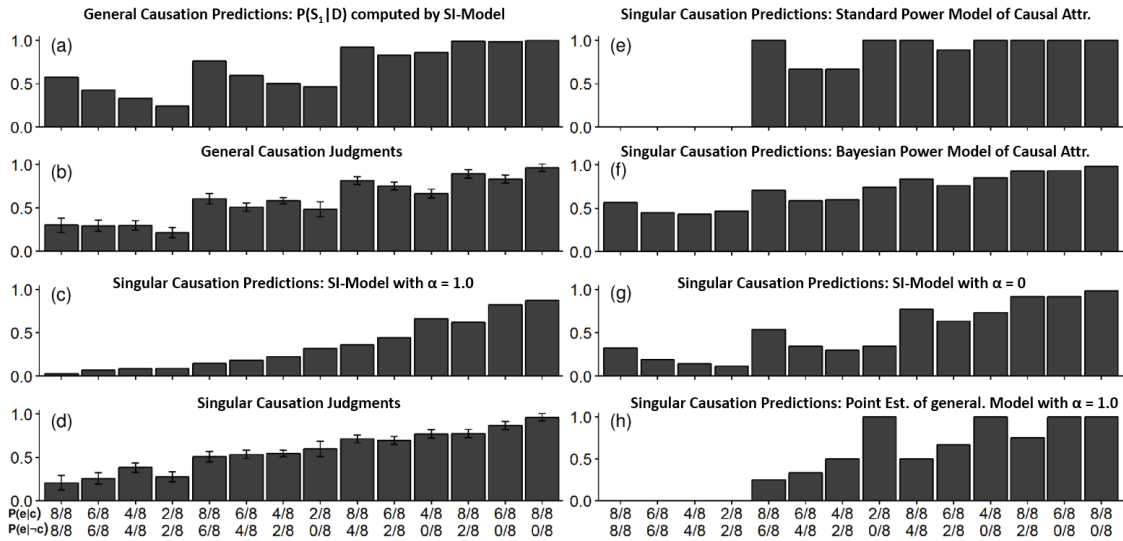


Figure 6.2: Predictions and results (means and 95% *CI*) of Experiment 2 (Stephan & Waldmann, 2018). Graph (a) shows the posterior probability of general causal structure S_1 and Graph (b) shows subjects' general causation ratings. Graph (c) shows the singular causation predictions made by the SI-Model using the generalized attribution equation with α set 1.0. Graph (d) shows subjects' singular causation ratings. The four graphs on the right show singular causation predictions made by different alternative models. Graph (e) shows the predictions by Cheng and Novick's (2005) standard model. Graph (f) shows the predictions made by the Bayesian variant of the model proposed by Holyoak et al. (2010) that incorporates parameter uncertainty. The predictions shown in Graph (g) use Cheng and Novick's standard equation but incorporate structure and parameter uncertainty. Graph (h) shows point estimations of the generalized attribution equation introduced in Chapter 4.

lar case to which the test question referred was randomly determined prior to the experiment and was kept constant for all subjects. Like in Experiment 1, subjects indicated how strongly they believed that the expression of the gene in the target mouse was caused by the respective chemical substance.

The mean ratings (based on $N = 82$ subjects recruited from an online panel) for the general and singular causation judgments are depicted in Graphs (b) and (d), respectively, in Figure 6.2. We found that subjects' general causation ratings (Graph b) were overall predicted well by the posterior probability of S_1 computed by the Structure Induction Model (Graph a). The results indicated that subjects' certainty as to whether the cause can generally produce the effect varied across the different data sets. This finding was important because we could otherwise not have argued that uncertainty about the general causal relationship has an impact on singular causation judgments. As predicted, we found that general and singular causation judgments were related but not equal. Generally, subjects tended to give higher singular causation judgments the more confident they were that the cause

can generally lead to the effect, but they also seemed to consider the possibility of causal preemption.

Most importantly, we found that the singular causation judgments (Graph d) were predicted well by our Structure Induction Model that computes $P(c \rightarrow e|c, e)$ with our generalized attribution equation and α set to 1 (Graph c). It can be seen in Figure 6.2 that our model predicts an overall smooth increase in the singular causation judgments from the first to the last data set. The mean singular causation ratings followed this pattern, even though they were on average slightly higher than the predictions. Figure 6.2 shows that all other models that we considered predicted a qualitatively different pattern of singular causation judgments. Different model fit measures (see Table 1 in Appendix A) that we computed confirmed that our model captured the results best. It yielded the overall highest correlation between predictions and mean ratings ($r = 0.94$) and it accounted for more variance in the observed ratings than all the other models ($R^2 = 0.88$). However, even though our model was superior on these global fit measures, the other models yielded quite high values here, too. The reason is that all models are sensitive to the contingency in the data. It was therefore important to also evaluate the models' performances for the subsets of the contingency set in which ΔP was kept constant. A particular strength of our model was that it also quite accurately accounted for these local trends, whereas all other models performed rather badly here. For example, the Bayesian variant of Cheng and Novick's attribution model (Graph f) that was proposed by Holyoak et al. (2010) predicted a negative trend for the first part of the contingency set where $\Delta P = 0$. The results, however, indicated a positive trend. Moreover, this model predicted local u-shaped trends that were clearly not observed. The Structure Induction version Cheng and Novick's model (Graph g) that neglects the possibility of causal preemption predicted negative trends for $\Delta P = 0$ and for $\Delta P = 0.25$. A particular problem of the standard attribution model and the generalized model, which both solely rely on point estimates (Graphs f and h), was that they could not handle the different situations in which $\Delta P = 0$. The reason is that their equations are undefined in this case. Furthermore, these two models predicted local trends that were qualitatively very different from the observed results. Overall, the results of Experiment 2 in Stephan and Waldmann (2018) indicated that subjects' singular causation judgments incorporated the possibility of causal preemption and that they were influenced by uncertainty about the existence of a general causal relationship between the target cause factor and the effect.

6.1.3 Summary

The goal of the studies in Stephan and Waldmann (2018) was to provide a first test of our new model and to demonstrate the principal relevance of both causal preemption and uncertainty about the general causal link between target cause and target effect for singular causation judgments. We tested our new model in the context of classical causal inductions tasks in which only statistical information was presented and temporal information was not provided. The studies focused on this type of context because it is the one that was originally targeted by the alternative models (Cheng & Novick, 2005; Holyoak et al., 2010). Also, we wanted to show that reasoners take the possibility of preemption into account even if temporal information, which together with causal power knowledge is supposed to be the type of information that determines intuitions about preemption, is not explicitly provided. The results of the experiments indicated that this was the case.

6.2 Experiments of Stephan, Mayrhofer, and Waldmann (submitted)

The studies of Stephan et al. (submitted) followed up on the experiments reported in Stephan and Waldmann (2018). Stephan and Waldmann (2018) demonstrated the principal relevance of causal preemption for singular causation judgments, but they did not experimentally manipulate temporal information. The goal of the studies reported in Stephan et al. (submitted) therefore was to test the different temporal factors introduced in Chapter 4 that determine α . Furthermore, as it was suggested in Chapter 4 that the probability of causal preemption of the target cause by alternative causes is given by the product of w_C , b_A , w_A , and α (Equation 4.1), another goal was to test the interplay of temporal and causal power information on singular causation judgments. Importantly, as the experiments aimed at evaluating specific predictions made by the generalized attribution equation (Equation 4.1), we tried to rule out any influence of general causal uncertainty. To rule out uncertainty about the general causal structure, a cover story was used in which it was explicitly mentioned that the effect can be brought about by each of two different causes². To minimize uncertainty about the size of the causal power parameters, the two potential causes were presented in two separate learning phases in which a relatively large number of observations was shown. Stephan et al. (submitted) conducted four

²Focusing on exactly one observed alternative cause A of the effect requires to set the base rate parameter b_A to 1, computing $P(c \rightarrow e|c, a, e)$ instead of $P(c \rightarrow e|c, e)$. The general computational principle of the model does not change, however

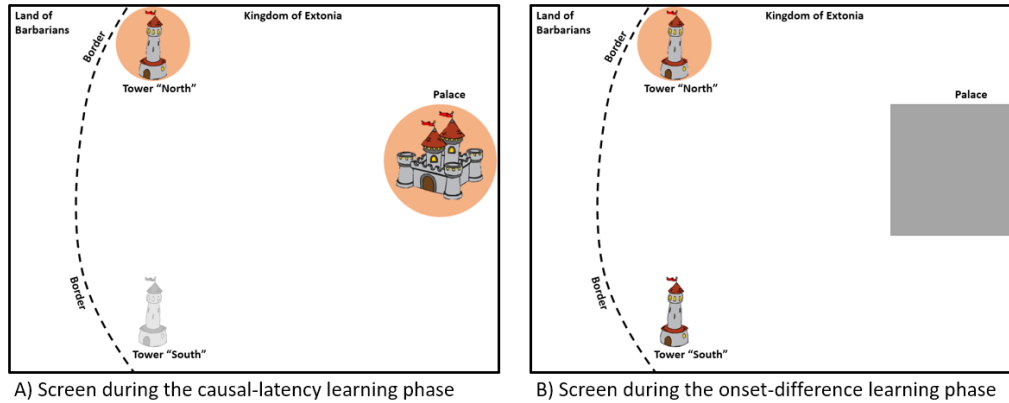


Figure 6.3: Illustration of the learning task used in Experiment 1 in Stephan et al. (submitted).

experiments in which different predictions made by the new generalized model were tested.

6.2.1 Overview of cover story, procedure, and materials

All four experiments used the same cover story and similar procedures. The cover story described a medieval kingdom, “Extonia”, that was threatened from time to time by hordes of vicious barbarians who tried to invade and plunder the empire. It was mentioned that the king, to protect his kingdom, had two watchtowers (tower “North” and tower “South”) built at the realm’s border. The tower crews were instructed to alarm Extonia’s knights by sending a carrier pigeon to the palace as soon as approaching barbarians were spotted. It was described that pigeons automatically set off an alarm the moment they arrived at the palace. The two watchtowers represented the two cause factors (C and A) and the alarm occurring in the palace represented the target effect factor (E).

In all experiments, subjects were asked to take the perspective of the king’s minister of defense who wants to inspect the efficacy of the empire’s defense system. Across the different experiments, different parameters were described to be relevant. In Experiment 1, for example, it was mentioned that the minister wants to inspect two different factors: the flight durations of each tower’s pigeons and the alertness of each tower, that is, how quickly each tower tends to spot approaching barbarians and sends off a pigeon. The first factor represented the causes’ causal latencies and the latter factor was supposed to capture differences in the onset times of the two causes. Subjects in all experiments were informed that they will make several observations of each tower in separate learning phases and that their task will be to learn about the different parameters. Before subjects proceeded to the learning

phase, it was mentioned that the final test question will be a causal query that refers to a singular occasion on which the palace was alarmed. We also presented several instruction-check questions that subjects had to answer correctly before they could proceed.

In all experiments, subjects learned the relevant parameters in dynamic animations. An illustration of the animations that were presented in the two learning tasks of Experiment 1 is shown in Figure 6.3. The left picture shows an illustration of the causal-latency learning task and the right picture shows an illustration of the onset-difference learning task. In the causal-latency learning task, the moment a tower sent a carrier pigeon was indicated by an orange circle occurring around tower. The arrival of the pigeon at the palace was signaled by a circle occurring around the palace. Subjects observed the two towers separately in the latency-learning task. They read that only the crew of one tower was allowed to send pigeons at a time, while the other tower was instructed to remain inactive. Subjects' task was to learn whether one tower had faster pigeons than the other. Subjects made repeated observations for each tower (the number of observations differed across experiments). In the onset-difference learning task both towers sent their pigeons on the same trial (as soon as they spotted the barbarians) and subjects' task was to learn whether there were systematic onset differences between the towers. Subjects made repeated observations here, too. As we wanted subjects to focus on the towers' onsets in this learning task and not on the causal latencies, the palace was masked by a grey rectangle. Several factors were counterbalanced in our experiments, including the order of the learning tasks, whether subjects first observed tower "North" or "South", which tower had the faster pigeons, which tower tended to spot barbarians earlier, what the target tower was to which the test question referred, and the orientation of the rating scale. A detailed description of all factors that were counterbalanced can be found in the methods sections of the original manuscript in Appendix C. After subjects had learned the respective parameters, they were presented the target scenario. The target situation described a singular occasion on which both towers sent a pigeon and on which an alarm occurred in the palace. Subjects read that the people of Extonia wanted to decorate the tower crew who caused the alarm on this occasion, but that there was the problem that both towers had sent pigeons. Subjects were asked to indicate how strongly they believed the alarm was caused by tower "North" ["South"] on this occasion. Whether the question referred to tower "North" or "South" was counterbalanced between subjects. We did not use a bipolar rating scale with the two towers on opposite sides of the scale because this would have created a disadvantage of Cheng and Novick's standard model, according to

which both potential cause factors should be considered as singular causes if they were simultaneously sufficiently powerful.

6.2.2 Experiment 1

Experiment 1 tested different scenarios in which the temporal factors were manipulated that determine the α parameter of our generalized model. As was shown in Chapter 4, these factors are the cause factors' causal latency and the difference in the causes' onset time. We have seen in Chapter 4 that the estimation of α requires the integration of both factors and that causal latency and onset difference are compensatory. We wanted to test whether reasoners indeed integrate the two factors as predicted by the model. As Experiment 1 was intended to serve as a first test, we used fixed values for the causal latencies instead of gamma distributions. In the causal-latency learning phase, subjects observed ten latencies per tower (20 observations in total). In the onset-difference learning phase, subjects made ten observations. The different parameter values for causal latency and onset difference can be found in Table 1 in Appendix C.

Four different scenarios were tested, which differed with respect to the combination of causal latency and onset difference (see Table 1 in Appendix B). In the first scenario, the two causes had identical causal latencies but differed with respect to their onset times. That is, subjects learned that one tower always spotted the barbarians earlier than the other tower. We have seen in Chapter 4 that in this case our model predicts that subjects should favor the cause factor that has the onset advantage on its side. The second scenario tested the complementary case in which the two towers had identical onset times but one tower had faster pigeons than the other. We here expected to see higher ratings when subjects were asked about the tower that had the faster pigeons. Scenarios 3 and 4 were the critical conditions, as we here pitted onset-time difference and causal latency against each other. This was done in two different ways that allowed us to test whether subjects actually integrated both temporal components or whether they considered one component to be more important than the other. Generally, our model predicts that subjects should give higher singular causation ratings for the cause that minimizes the sum of causal latency and onset time (see Equation 4.2). In Scenario 3, one cause factor was instantiated repeatedly after 1600 *ms* and had a causal latency of 800 *ms*, while the other cause always occurred after 800 *ms* and had a causal latency of 2400 *ms*. Thus, the first cause minimized the sum of onset time and causal latency (2400 *ms* vs. 3200 *ms*) and thus was expected to receive higher singular causation ratings. However, higher singular causation ratings for this cause could have been due to

an unconditional preference for shorter causal latencies. Scenario 4 controlled for this confound by testing the opposite case. Here it was the cause with the onset advantage but the causal-latency disadvantage that minimized sum of onset time and latency.

The causal powers of the two causes were not manipulated in this first experiment because we wanted to focus on the temporal components identified by our model. Both causes had a causal power of 1. That is, subjects observed that each tower always succeeded in setting off an alarm in the palace. Cheng and Novick's standard model hence predicted very high and identical ratings across all four scenarios, while our model predicted the differences that were described above.

The results (based on $N = 384$ subjects recruited from an online panel) of Experiment 1 were in line with the predictions of our model, indicating that subjects not only incorporated information about onset-difference and causal latency but that they indeed integrated the two types of information in a compensatory manner. A graphical summary of the results can be found in Figure 8 in Appendix C. In the first scenario, subjects gave higher ratings when they were asked about the cause that always occurred first than when asked about the cause that always occurred second, following the principle that a cause that occurs earlier is less likely to be preempted by a competitor if everything else can be assumed to be constant. In Scenario 2, in which the onset-difference was kept constant, subjects gave higher singular causation ratings when asked about the cause with the smaller causal latency, following the principle that, everything else being equal, faster causes are less likely to be preempted. Importantly, the ratings in both Scenario 3 and Scenario 4 were higher for the cause factor that minimized the sum of causal latency and onset time, no matter whether the advantageous factor was causal latency or onset time. Moreover, we found that the differences in the ratings for the two potential causes were almost identical in Scenarios 3 and 4, which additionally corroborated that subjects considered both factors to be equally important. However, we also found that the differences between the causes were smaller than in Scenarios 1 and 2. The fact that the observed differences were overall smaller in Scenarios 3 and 4 than in Scenarios 1 and 2, where the causes differed on only one temporal dimension, can be explained by the higher difficulty of the computation in the more complex cases. The more complex conditions were also more memory demanding. Subjects here had to retrieve from memory information about the degree to which the causes differed on both dimensions, while it sufficed in the first two scenarios to remember the ordinal ranking of the causes on a single discriminating dimension. Overall, the results of Experiment 1 suggested that subjects indeed estimated the probability of

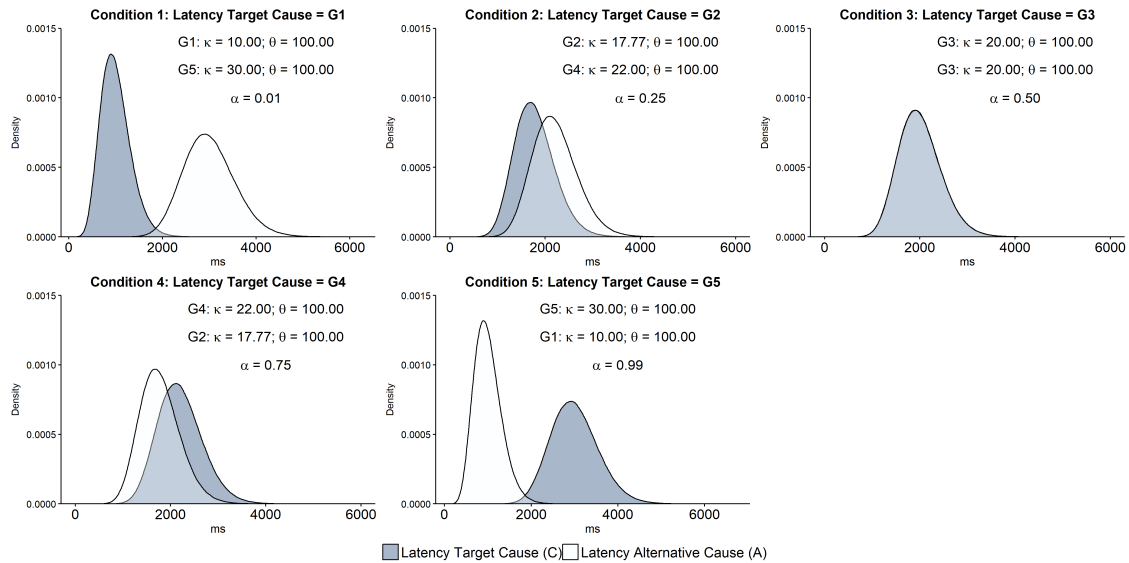


Figure 6.4: Pairs of gamma distributions contrasted in the five conditions of Exp. 2 in Stephan et al. (submitted). The shape (κ) and scale (θ) parameters of the five different gamma distributions are listed for each pair. The depicted α values were obtained using the MC algorithm corresponding to Eq. 4.2. The dark distributions show the causal latencies of the target cause and the light distributions show the causal latencies of the alternative cause.

causal preemption as is predicted by our Equation 4.2.

6.2.3 Experiment 2

The goal of Experiment 2 was to test the more natural case of variable causal latencies. We compared different conditions in which the two potential causes were paired with causal latencies that followed different gamma distributions. On the assumption that variation in causal latencies is the default situation, we expected subjects to be sensitive to this factor and hypothesized that their singular causation judgments are in line with the predictions of our generalized model. However, another possibility that we considered was that subjects might only estimate the expected values of the causal latencies, while variability in causal latency is neglected. The way in which we manipulated causal latencies in Experiment 2 allowed us to test both possibilities. As we wanted to focus on causal latency information in this study, we again restricted the scenarios to deterministic causes. Furthermore, as we wanted to keep the onset-difference constant, we informed our subjects that both causes occurred simultaneously in the target situation.

The pairs of gamma distributions that were contrasted in five different between-subjects conditions are shown in Figure 6.4. As can be seen, the “distances” between the distributions (G1 - G5; higher numbers indicate higher expected values) in

the different conditions varied quantitatively, from fairly extreme in Condition 1, for example, to identical in Condition 3. The latency distribution belonging to the target cause C in each condition is depicted in dark blue. In Condition 1, for example, which contrasted distributions G1 and G5, the target cause's latency followed G1, whereas the latency of the alternative cause followed G5. Conditions 4 and 5 involved the same pairs of distributions as Conditions 1 and 2, but the distribution associated with the target cause was changed. Figure 6.4 also shows the different α values estimated with the sampling algorithm introduced in Chapter 4. For the first pair, for example, α equals 0.01. In the case of deterministic causes ($w_C = w_A = 1$), α directly corresponds to the probability that c was preempted by a . Thus, we expected participants to be confident in this condition that it was indeed the target cause c that brought about the observed outcome. In the fifth condition, by contrast, in which C follows G5 and A follows G1, participants were expected to be confident that c did not cause the effect. Figure 6.4 also shows that all the other conditions were expected to elicit more uncertainty concerning the preemptive relation. In the third condition, for example, in which C and A had the same latency distribution (G3), which means that $\alpha = 0.50$, we expected participants to be uncertain about whether tower "North" or tower "South" caused the effect. In the case of deterministic causes, the predictions for $P(c \rightarrow e|c, a, e)$ are given by $1 - \alpha$. Our model thus predicted a negative linear trend of the ratings, with the highest ratings in Condition 1 and the lowest ratings in Condition 5. The version of our model that computes α based on only the expected values of the latency distributions predicted maximal ratings for Conditions 1 to 3 (Condition 3 is represented as a situation of symmetric overdetermination in this case) and zero ratings for Conditions 4 and 5. The standard model predicted maximal ratings in all conditions. The predictions of the different models are shown in Figure 6.5.

Other than in Experiment 1, the learning phase consisted of only the causal-latency task (see Figure 6.3 A). Subjects were shown thirteen observations per tower, which corresponded to thirteen quantiles of the respective latency distribution. We used quantiles instead of randomly sampled latencies because we wanted to keep the overall sample of observations relatively small but representative for the respective latency distribution. The thirteen causal latencies per tower were presented in random order. The target situation to which the singular causation test query referred was similar to the one in Experiment 1, with the exception that subjects were informed that it was known that both towers sent their pigeons simultaneously (i.e., $\Delta_t = 0$ and could be neglected).

The results (based on $N = 200$ subjects recruited from an online panel) were in

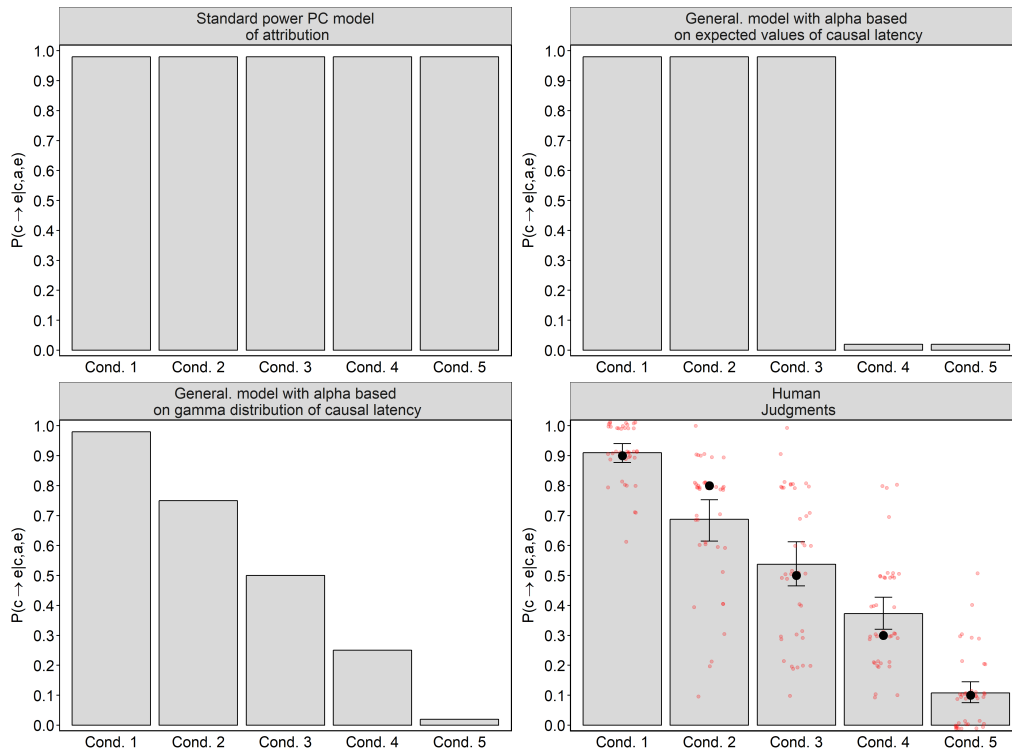


Figure 6.5: Model predictions and results of Experiment 2 in Stephan et al. (submitted) (bars show mean singular causation ratings; error bars denote 95% bootstrapped *CIs*; red points show jittered individual ratings; black points show medians) for the different conditions (see Fig. 6.4). The predictions of the Standard power PC model of causal attribution were obtained from Equation 3.2. The predictions of the generalized model were based on Equation 4.1 and different ways of estimating α . The first of these plots shows the predictions obtained when α is calculated based on the expected values of the respective gamma distributions. The second plot shows the predictions obtained when α is estimated based on the gamma distributions.

line with the predictions made by our generalized model that uses the full causal-latency distribution to compute α . The results were at odds with both Cheng and Novick’s (2005) standard model and the version of our model that relies on only the expected values of the causal latencies (see Figure 6.5). Subjects’ singular causation judgments were not only sensitive to causal latency but also took the variability of causal latency into account.

6.2.4 Experiment 3

In Experiments 1 and 2 deterministic causes were tested because we aimed to isolate the influence that the introduced temporal factors exert on singular causation judgments. However, according to our generalized model (Equation 4.1), causal power information should also play a crucial role in singular causation judgments. A cen-

tral new prediction of our model is that causal power and temporal information are expected to interact, as the probability of preemption is given by the product of w_C , w_A , and α . The goal of Experiment 3 was to explore this core property of our generalized model.

A scenario was tested in which both causes either had a causal power of $w_C = w_A = 0.83$ or of $w_C = w_A = 0.5$. Causal latency was manipulated by using the first pair of gamma distributions (G1 vs. G5) shown in Figure 6.4. The causal latency of the target cause was either associated with G1 or G5, while the causal latency of the alternative cause always followed the complementary distribution of the pair. This combination of causal powers and relative causal latencies of the potential causes led to a predicted interaction effect that is shown in the middle panel of Figure 6.6. The x-axis shows the relative causal latency of the target cause, that is, whether it followed distribution G1 or G5. As can be seen, when the target cause's relative causal latency is high (i.e., its causal latency follows G5 while the causal latency of the alternative cause follows G1), our generalized model predicts that ratings should be higher when the causal power of the target cause is low than when it is high (black vs. grey bars in Fig. 6.6). Conceptually, this condition represents a scenario in which a reliably working target cause is competing with an alternative cause that not only is also working reliably but that also tends to operate quicker than the target cause. This should make it very unlikely that the target cause was the singular cause of the effect on an occasion on which both factors were present. Mathematically, the product that is subtracted from w_C in the high-power condition (black bar in the left pair) is sufficiently large so that the numerator ends up smaller ($0.83 - 0.83 \cdot 0.83 \cdot 0.99 = 0.15$) than in the low-power condition (grey bar in the left pair; $0.5 - 0.5 \cdot 0.5 \cdot 0.99 = 0.25$). Furthermore, the denominator is smaller in the low-power condition than in the high-power condition ($0.5 + 0.5 - 0.5 \cdot 0.5 = 0.75$ vs. $0.83 + 0.83 - 0.83 \cdot 0.83 = 0.97$), which means that the numerator in the low-power condition is increased more strongly than in the high-power condition. Figure 6.6 also shows that the new model predicts the reversed effect when the relative causal latency of the target cause is low (i.e., when it follows G1; see right pair of bars in the middle panel of Fig. 6.6). Here, the product that is subtracted in the numerator from the target cause's causal power becomes small because the target cause is operating quicker than its competitor. The causal-latency advantage that the target cause has allows it to manifest its high causal power. In this condition, our model makes the same predictions as the standard power PC model of causal attribution: we expected subjects to give higher singular causation ratings in the high causal power condition than in the low causal power condition. Finally, Figure 6.6 shows that

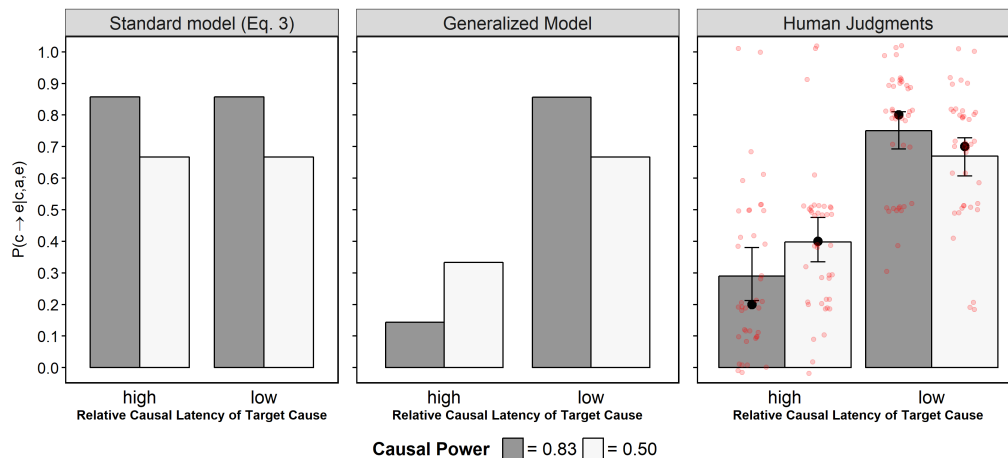


Figure 6.6: Model predictions and results of Exp. 3 in Stephan et al. (submitted) (bars represent mean singular causation ratings; error bars denote 95% bootstrapped *CIs*; red points show jittered individual ratings; black points show medians). The x-axis shows the relative causal latency of the target cause.

the new model also predicts a main effect of causal latency: causes that on average tend to precede the efficacy of their competitors (right pair of bars) should receive higher singular causation ratings than causes whose competitors tend to preempt them (left pair of bars). Finally, a main effect of causal power was predicted by the standard but not by our generalized model.

Causal power and causal latency were manipulated between subjects in this experiment. The materials and procedure were largely identical to those of Experiments 1 and 2, with the exception that information about the probabilistic causal nature of the towers was added to the instructions. Participants read that pigeons might get lost on their way to the palace and that it would therefore be important to study the pigeons' arrival rates. Secondly, the learning task was modified accordingly, so that causal power and causal latency information were conveyed together in a single learning phase. Other than in Experiment 2, subjects observed 24 pigeons per tower. The number of observations was increased to ensure that all participants observe a sufficient number of "successful" pigeons to be able to learn the towers' causal latencies. Subjects in the high-power condition observed 20 successful pigeons per tower, whereas subjects in the low-power condition observed 12 successful pigeons per tower. The flight durations corresponded to 12 or 20 quantiles of the respective causal latency distributions. Whenever a pigeon failed to reach the palace during the learning task, the words "pigeon probably lost" were displayed after a duration that corresponded to the 99.9th percentile of the "slower" causal-latency distribution (G5). The description of the target situation and the singular causation test question was identical with the one in Experiment 2. Additionally, on a

separate screen, subjects were asked to estimate the causal powers of the two causes. We assessed causal power ratings because we wanted to control for the possibility that subjects' representations of the causal powers of the two causes might have been influenced by the causes' causal latencies. Our generalized model considers causal power and causal latency to be two independent properties of causes: While causal power solely reflects causal strength (cf. Cheng, 1997), causal latency solely reflects the time it takes a cause to generate its effect. Yet, we wanted to control for the possibility that subjects might not have kept causal power and latency information apart, and instead represented a relatively slow cause as less powerful than a relatively fast cause. This would be problematic for our generalized model because the predicted main effect of causal latency could then be explained by perceived differences in the causal powers alone.

The results (based on $N = 160$ subjects recruited from an online panel) of the experiment were in line with the predictions of our generalized model (see Figure 6.6). First of all, we found the predicted main effect for causal latency. Subjects tended to give higher ratings when the target cause had a smaller causal latency than its competitor. This finding replicated the previous results from Experiments 1 and 2 that indicated that reasoners' singular causation judgments are sensitive to causal latency information. Secondly, the main effect of causal power that was predicted for the tested combination of parameters by the standard model but not by our generalized model was not found. Most importantly, the singular causation judgments showed the interaction effect between causal power and temporal information that was predicted by our generalized model. This finding suggested that subjects indeed integrated temporal information about causal latency (captured by α) and information about the causal powers of the potential causes (w_C and w_A) as predicted by our generalized model. However, what can also be seen in Figure 6.6 is that the interaction effect was smaller than predicted.

Finally, the ratings for the causal-power control questions (see Figure 13 in Appendix B) did not indicate an influence of causal latency on causal-power representations. Subjects in the different causal-latency conditions gave causal-power ratings that were close to the normative values (of 0.50 and 0.83). The observed main effect of causal latency was thus not due to distorted representations of the cause factors causal powers.

6.2.5 Experiment 4

Although the pattern of singular causation judgments in Experiment 3 was captured well by the new model, the predicted interaction effect was relatively weak.

As the predicted interaction effect follows from a core principal of our generalized model, we decided to replicate the results in a fourth experiment that investigated a set of causal power parameters that should lead to a larger interaction effect. The same causal latency distributions were used (distributions G1 and G5), but this time causes were contrasted that either had a power of $w_C = w_A = 0.93$ or $w_C = w_A = 0.40$. The model predictions are depicted in Figure 6.7. Another difference from Experiment 3 was the control question that we asked in the end. The control question that we asked at the end of Experiment 3 only controlled for the possibility that subjects' causal power representations were distorted by causal latency information. However, an influence in the other direction is also possible, that is, differences in the causal powers might have an impact on causal latency representations. We assumed that this is not unlikely in the paradigm that we used in our studies. The trials in which a cause failed to generate the effect ended after a duration that corresponded to the 99.9th percentile of the slower causal latency distribution. Consequently, subjects in the low-causal power condition spent more time on the task than subjects in the high-causal power condition. At the end of Experiment 4, we therefore asked a control question that assessed how strongly subjects' causal latency representations are affected by the causal power information. Subjects were asked to imagine a situation in which both towers sent their pigeons simultaneously, and then had to indicate on a rating scale which pigeon they believed to be quicker (i.e., which tower had the smaller causal latency). Answers to this question also allowed us to calculate the α value of our model based on subjects' representations.

The results of Experiment 4 (based on $N = 384$ subjects recruited from an online panel) successfully replicated those of Experiment 3 (see Figure 6.7). We obtained again a significant main effect for causal latency and, crucially, also a significant interaction between causal power and causal latency. Moreover, the bigger differences in the causal power parameters that we used (0.93 vs. 0.40 in Experiment 4 compared to 0.83 vs. 0.50 in Experiment 3) did indeed lead to a larger interaction effect. However, it can be seen in Figure 6.7 that the differences were still smaller than predicted by our model.

The ratings for the causal latency control question were overall high in both causal-power conditions, confirming that subjects correctly identified the cause that had the smaller causal latency (an illustration is shown in Figure 15 in Appendix C). Yet, we also found that the causal power difference led indeed to a small distortion of subjects causal-latency representations. When the causal power of the two causes was high, subjects were more likely to give a rating associated with the cause with

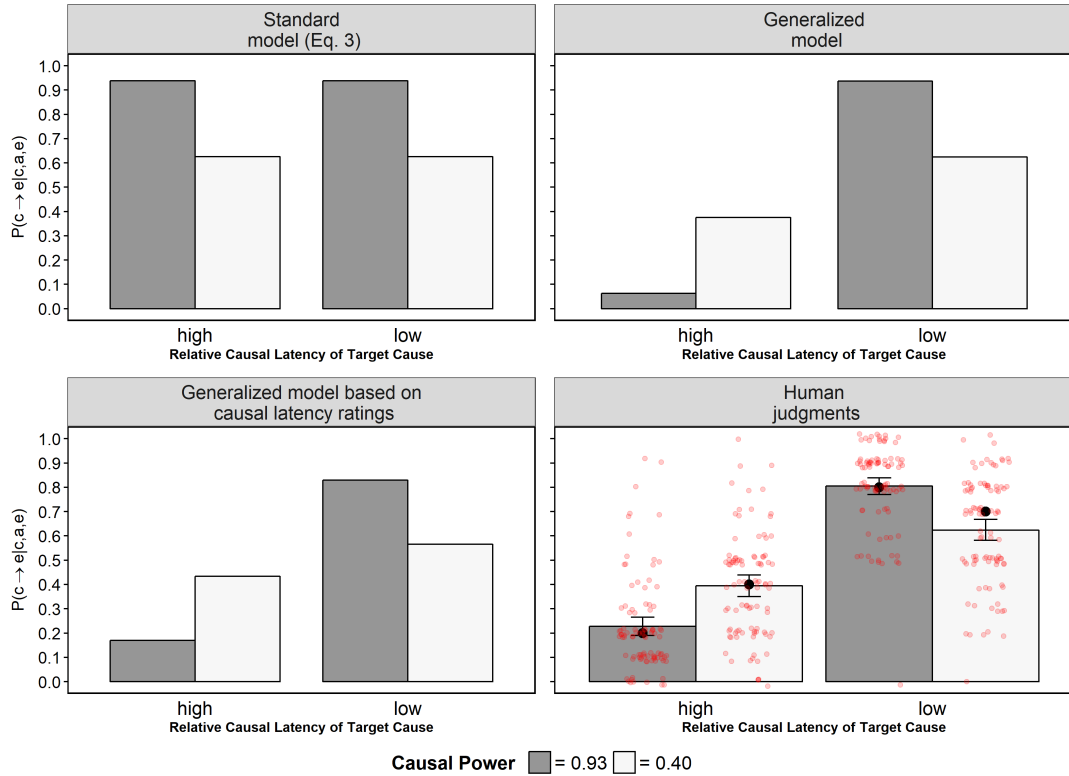


Figure 6.7: Model predictions and results of Experiment 4 in Stephan et al. (submitted) (bars represent mean singular causation ratings; error bars denote 95% bootstrapped *CI*s; red points show jittered individual ratings; black points show medians). The x-axis shows the relative causal latency of the target cause.

the lower causal latency than when the causal power of the causes was low.

We used subjects' causal latency ratings to estimate the α parameter in our model, to see how this would change the predictions for the singular causation ratings. The model predictions based on subjects' causal latency ratings are depicted in the third panel in Figure 6.7. It can be seen that this model did indeed predict a smaller interaction effect than the model that used the objective parameter values. The results of this experiment demonstrated again the validity of the new model.

6.2.6 Summary

While the studies of Stephan and Waldmann (2018) did not manipulate the temporal factors that according to our generalized model are crucial for judgments of singular causation, the studies of Stephan et al. (submitted) tested central predictions concerning the interplay of causal power and temporal information. Experiment 1 demonstrated that reasoners are sensitive to both onset difference and causal latency, and that they actually seem to integrate the two temporal dimensions in a

compensatory manner, as predicted by our model (see Equation 4.2). Experiment 2 demonstrated that reasoners' singular causation judgments are sensitive to variability in the potential causes' causal latencies. Finally, Experiments 3 and 4 tested the crucial prediction of the new model that causal power and temporal information should interact. These last two experiments are particularly important, because we here varied all parameters that are part of the new generalized power model of causal attribution. The results supported the model.

Chapter 7

General Discussion

This thesis asked how causal queries about singular cases can be answered, that is, how it can be assessed what the cause of an observed effect e was. Cheng and Novick's (2005) Power Model of Causal Attribution, which belongs to the class of normative computational models and relies on Cheng's (1997) famous Causal Power Theory, was the focus of this thesis. According to the Power Model of Causal Attribution, the relevant information for the assessment of singular causation is the relative strength or causal power of the potential causes of the observed effect¹. The core computational principal behind the model, as I have shown, is to determine the probability with which a target cause factor C was sufficiently powerful on an occasion to produce the effect. This dissertation pursued the goal of demonstrating that more needs to be considered to answer singular causation queries. First of all, I showed that a successful model also needs to take temporal information (Chapter 4) into account. The reason is that it can otherwise not handle the problem of causal preemption. Also, in Chapter 5 I argued that uncertainty about the general causal structure and the causal parameters needs to be considered. I have proposed a new model that generalizes the standard equations so that they can handle the problem of preemption. I have then proposed a Bayesian extension of the model that incorporates general-causal uncertainty. The experiments that were conducted demonstrated the validity of the new model and confirmed several of its predictions. However, some of the model's aspects and predictions have remained yet untested. In the following, I first want to turn to and discuss these empirical limitations. I then will finish with an elaboration on more general theoretical issues that I consider to be important and that I think provide interesting avenues for future research projects.

¹We have seen that in situations in which it is not known whether the potential causes of the effect were actually present, their base rates are also regarded as relevant.

7.1 Empirical limitations

A crucial prediction of the Structure Induction Model of Singular Causation Judgments that was presented in Chapter 5 is that reasoners should be sensitive to the sample size of observations, as a general principle that is formally captured by any Bayesian updating model is that the confidence in a particular hypothesis should increase the more data one has observed. A shortcoming of the study of Stephan and Waldmann (2018) is that no experiments were conducted that directly tested this prediction. The sample size varied between Experiment 1 and 2 but a direct test was not conducted. A comparison of the singular causation ratings for the same contingencies across the two experiments reported in the study seems to suggest that subjects were at best mildly sensitive to sample size. Yet, the two experiments varied with respect to the experimental design and are thus not fully comparable. It would thus be interesting to manipulate this factor in a future experiment.

Another limitation of the study by Stephan and Waldmann (2018) was that only one type of cover story was used in Experiment 2, where the influence of general-causal uncertainty on singular causation judgments was investigated. The Structure Induction Model of Singular Causation Judgments, following previous research by Griffiths and Tenenbaum (2005) and Meder et al. (2014), regards maximal prior uncertainty as the default, which is modeled with uninformative priors over the causal structure and the parameters. In the chemical-gene scenario that was tested, maximal uncertainty about the general causal structure and the parameters seemed to be a plausible assumption, but in other contexts this might well be different. In fact, Lu et al. (2008) have argued that reasoners' causal inferences would often be in line with a Bayesian learning model that assigns a *strong and sparse* prior to the causal strength parameters (see Griffiths, 2017, for an overview). Lu et al. (2008) have not investigated singular causation judgments, however, and it would thus be interesting to test our model under different types of scenarios.

One noteworthy characteristic of the experiments in Stephan et al. (submitted), which manipulated the temporal factors that are incorporated by the Generalized Power Model of Causal Attribution, was that subjects observed all potential causes of the effect. In most real-world situations, however, reasoners observe only a subset of the potential causes and thus typically have to deal with the problem that the alternative causes are unobservable. The experiments in this study have not considered situations with unobserved background causes because this would have required a substantial modification of the paradigm. However, I have outlined in Chapter 4 that the Generalized Power Model of Causal Attribution can handle this type of situation. It was suggested there that the temporal distribution of the base rate of

the effect can be modeled with exponential distributions. The observed causal-delay distribution in the presence of the target cause would then be a mixture of the particular gamma distribution of the target cause’s causal latency and the exponential background distribution, but the same algorithm for the calculation of α could be used to derive the predictions. It would be interesting to test such situation in future experiments and to examine whether the model accurately captures people’s judgments here, too.

A common feature of the experiments in Stephan and Waldmann (2018) and Stephan et al. (submitted) is that only static target situations were tested. In all experiments, subjects were asked to make singular causation judgments based on previously acquired knowledge about causal strength, causal delays and onsets differences. The structure of the target situation that was tested thus resembled a classical crime scenario in which the detective enters the scenery and tries to identify the perpetrator after all events had unfolded. The reason why the studies focused on this type of test situation was that it is the standard situation analyzed by the Standard Power Model of Causal Attribution and because it is the type of situation that was tested in previous related studies (e.g., Holyoak et al., 2010). In many real-life situations, however, singular causation queries arise in dynamic contexts. A crucial feature of dynamic test cases is that the onsets of the relevant events and the delays between them are experienced. What the presented model thus far neglects is that experiencing the singular delay between potential causes and effect should introduce a dependency between the causal power and temporal parameters of the model, as information about the delay between the onset of a target cause and the onset of the effect should be informative about the causes’ actual causal strengths in the situation. Imagine a target cause with an average causal strength of $w_C = 0.90$ and an expected causal delay of 20 minutes. Think of a person who takes a dose of Aspirin and who experiences relief from her pain only after two hours. In this situation, the unexpectedly long delay seems to provide evidence that the causal strength of the drug was zero on the given occasion, and it therefore should weaken the belief that it was the singular cause of the effect. It would be interesting to model and test such situations in future studies.

Another prediction made by the Power Model of Causal Attribution that was not tested in the reported experiments is that reasoners’ singular causation judgments should be sensitive to the base rate of the potential causes in situations in which only the effect is observed. It was shown in Chapter 3 that singular causation judgments are supposed to follow $P(c \rightarrow e|e)$ in this type of situation and that the numerator of the corresponding equation consists of the product of the base rate of the target

cause C , denoted b_C , and its causal power, w_C . This type of situation can be easily investigated in future experiments by augmenting the developed learning paradigm with a separate base-rate learning phase.

7.2 General theoretical considerations and outlook

7.2.1 The role of knowledge about causal mechanisms

An important question that this thesis did not address concerns the role of causal mechanism information in the assessment of singular causation. Causal mechanism information has, however, been identified as an important type of information in causal reasoning (an overview is given by Johnson & Ahn, 2017). It has been shown, for example, to play an important role in how reasoners form causal explanations (e.g., Lombrozo, 2010; Lombrozo & Vasilyeva, 2017; Nagel & Stephan, 2015, 2016) or make predictive judgments (e.g., Bes, Sloman, Lucas, & Raufaste, 2012; Stephan, Tentori, Pighin, & Waldmann, in preparation). Causal mechanism approaches to causality are also one of the central philosophical frameworks of causality (e.g., Glennan, 2017; Machamer, Darden, & Craver, 2000). What is of particular relevance for the project of this thesis is that different studies have shown that reasoners regard mechanism information as particularly relevant when making causal attributions (Ahn, Kalish, Medin, & Gelman, 1995; Johnson & Keil, 2018). Moreover, in a recent article about normative strategies for the assessment of singular causes, the philosopher Nancy Cartwright (2017) has listed mechanistic information among the factors that help assess singular causation. It is therefore important to say something about how mechanism information interrelates to the model that I have developed and defended in this thesis.

I think that the influence of mechanism information on singular causation judgments can be captured by the proposed Generalized Power Model of Causal Attribution by the addition of variables. Generally, causal mechanisms can be represented in causal graphical models by additional variables that are interpolated between the cause and the effect factors (an illustrative overview of how to formally capture and utilize causal mechanism information in causal reasoning can be found in Chapter 9 in Pearl & Mackenzie, 2018), thereby turning a previously *direct* causal connection (e.g., $C \rightarrow E \leftarrow A$) into an *indirect* one ($C \rightarrow M_C \rightarrow E \leftarrow M_A \leftarrow A$). For an illustration of how mechanism information can be incorporated by the model and to illustrate why mechanism information is valuable when it comes to the assessment of singular causation, let us consider the type of situation again in which a reasoner knows that C , A , and E occurred on a singular occasion and she is asked to express

how confident she is that c instead of a caused e on this occasion. Knowledge about the status of the mechanism variable M_A connecting A and E would be helpful here because knowing the status of M_A constrains a 's actual causal power on the target occasion. If the causal chain $A \rightarrow M_A \rightarrow E$ honors the causal Markov condition – which states that a variable in a causal network that is conditioned on its direct cause is independent from all other variables except for its own descendants – and if there is only one “mechanism path” leading from A to E (on which M_A lies), the absence of M_A ($\neg m_A$) on a singular occasion ($\neg m_A$) implies that a had a causal power of zero in this case and can therefore be ruled out as a singular cause of e . Consequently, whenever c is the only remaining potential cause of the effect, one can safely conclude that e was caused by c . I have mentioned in Chapter 3 that an important strategy for the assessment of singular causation is the elimination of alternative causes (“Holmesian inference”), and we have there considered an example in which an alternative cause A was found to be absent on the target occasion. Finding that a mechanism variable of an alternative cause A was not instantiated on the target occasion is another means of eliminating A from the list of the potential causes. The Power Model of Causal Attribution can be easily augmented to incorporate this mechanism-checking strategy. Consider how Equation 4.1 proposed in Chapter 4 would have to be modified to capture such a situation. What would change is that another conditional element would have to be added to $P(c \rightarrow e|c, a, e)$, turning this expression into $P(c \rightarrow e|c, a, \neg m_A, e)$. On the right-hand side of the equation, the unconditional causal power of A , w_A , would have to be substituted by A 's power conditional on the absence of M_A , $w_{A|\neg m_A}$. Assuming that the causal Markov condition holds and that there is only a single mechanism path connecting A and E , $w_{A|\neg m_A}$ would amount to zero and $P(c \rightarrow e|c, a, \neg m_A, e)$ would reduce to $\frac{w_C}{w_C} = 1$. Of course, even in cases in which A produces E via multiple mechanism paths, assessing the status of M_A can be helpful. If a situation can be conditionalized on the absence of M_A , this should generally increase $P(c \rightarrow e|c, a, \neg m_A, e)$, as the causal Markov condition implies that $w_{A|\neg m_A} < w_A$.

Importantly, causal mechanism information might not only constrain the causal strength parameters in the model but also the parameter α . The reason for this is that a causal mechanism connecting a target cause and a target effect might consist of multiple causal paths which might differ not only with respect to their causal powers but also with respect to their causal latencies. Take the example of a coroner who wants to find out whether a victim displaying an abdominal gunshot wound actually died because of the gunshot. The causal strength with which bullets kill their victims is surely quite high on average, but also fairly variable. Similarly,

bullets typically bring death to their victims relatively quickly, but not always. Depending on the organs that are damaged, both the probability of dying from a gunshot and the latency can be high or low. A case in which the bullet took a straight path to the victim's heart will probably lead to higher confidence in a singular causal link between gunshot and death than a case in which the bullet left all organs intact. Bullets that stop their victims' hearts are not only more effective, they also bring death so quickly that there is less room for alternative causes to strike (e.g., an assassin might first have administered a lethal dose of poison but then decided to shoot the victim). Manipulating and testing the influence of causal mechanism information on singular causation judgments will be an interesting avenue for future studies.

7.2.2 Assessing singular causation vs. selecting singular causes

The new Power Model of Causal Attribution proposed and evaluated in this thesis addresses the problem of how it can be determined whether an observed singular co-occurrence of events was actually causal ($c \rightarrow e$) instead of coincidental (c, e). A different problem discussed in the literature (e.g., Hitchcock & Knobe, 2009; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Phillips, Luguri, & Knobe, 2015) under the name singular or actual causation refers to the interesting phenomenon that even though several factors typically need to come together to generate an effect, people have the strong tendency to regard only a subset, or even only one, of these factors as *the* cause(s) of the observed outcome (see also Cheng & Novick, 1991; Novick & Cheng, 2004). For an illustration of this phenomenon, let us consider the frequently discussed forest fire example (e.g., Halpern & Hitchcock, 2015), in which it is stated that a forest fire occurred after a lit match had been dropped negligently. In this example, most people are more inclined to regard the negligently dropped lit match as the cause of the fire than the presence of oxygen in the surrounding atmosphere or than dry material on the ground. Although we would probably not deny causal relevance of these latter factors, they are regarded as mere conditions that enable the lit match to cause the effect.

Importantly, selecting the lit match over oxygen and dry material presupposes that the lit match had already been identified as a *causally relevant* factor on the singular occasion that is considered. This model proposed in this dissertation targeted the question of how the causal relevance of a potential cause on a singular occasion can be assessed and not what the factors are that determine whether a causally relevant event is selected over others. Just imagine that there also had

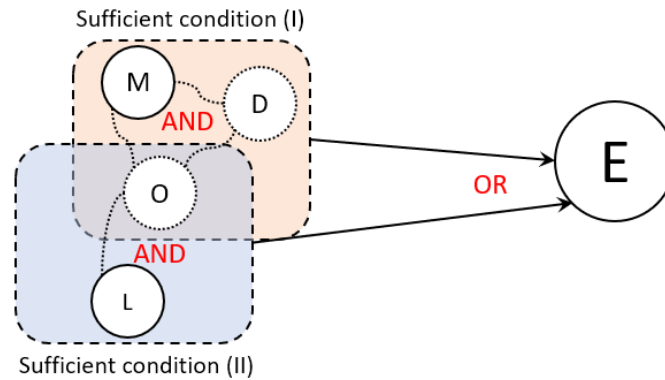


Figure 7.1: Illustration showing the difference between the concepts of singular causation and causal selection. The different cause factors sharing a colored area constitute a (probabilistically) sufficient condition for the effect. All cause factors within an area, by contrast, are conjunctively connected.

been a lightning strike on the day the forest fire occurred. This opens up the possibility that the lit match was completely irrelevant, and one could thus say that it only coincidentally co-occurred with the effect.

The difference between the two types of problems is illustrated in Figure 7.1, where the potential cause factors of a common effect factor are grouped together in different colored areas. Each area represents a non-redundant conglomeration of factors that are only jointly sufficient for the effect, though still only probabilistically sufficient unless the set comprises an exhaustive list of factors (cf. Mackie's, 1974, idea of INUS conditions). Importantly, all factors subsumed under a common area are conjunctively connected (formally expressed: they are connected by a logical "AND"; e.g., match *and* dry material *and* oxygen), whereas any two factors belonging to different areas represent potential *alternative* causes of the effect (logical "OR"; e.g., lit match [and oxygen and dry material] *or* lightning strike [and oxygen]). Singular causal selection refers to the highlighting of one of the factors within a probabilistically sufficient condition, whereas what I would like to call assessing singular causation, and which is targeted by the model evaluated in this thesis, refers to the process by which the effect is attributed to one of different alternative factors.

Establishing the causal relevance of a factor on a singular occasion epistemically comes before the causal selection process, but this does not imply that the selection of a factor over others is generally less important. The selection of a factor will, for example, often be of enormous *pragmatic* relevance. In the forest fire case, a reasonable goal would be to prevent a similar incident in the future. For this purpose it appears extremely relevant to highlight the negligently dropped lit match, whereas

the presence of oxygen is of no pragmatic value in this respect.

The goal of identifying good “target points of intervention” has indeed been described in the literature as a relevant factor behind the selection process (see, e.g., Hitchcock & Knobe, 2009). Another prominent idea in the literature is that the selection of singular causes be driven by considerations about their “normality” (Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986; Kominisky et al., 2015; Phillips et al., 2015). The idea is that those singular causes are preferably selected whose occurrence is more abnormal than that of others in the target situation. Interestingly, events that fulfill the abnormality criterion will often also be those that fulfill target of intervention criterion. “Abnormality”, under these accounts, can refer to both statistical normality or prescriptive norms (for a critical view on the role of prescriptive norms in causal selection see Samland & Waldmann, 2016).

An interesting future project might be to develop an integrative account that can handle both types of problems. An obvious approach would be to construct and test a model that runs through a two-step process. In a first step, the causally relevant factors are separated from those that were causally irrelevant in a given situation. Here, the model would consider the factors that were identified as relevant by the Generalized Power Model of Causal Attribution proposed in this thesis, that is, it would consider causal power and temporal information. In a second step, the model would take into consideration those factors that were identified as important for the selection process.

7.2.3 Singular events and their corresponding types

The Power Framework of Causal Attribution commits to the philosophical point of view that singular causal connections between occurred events are not directly observable, and that the assessment of singular causation therefore needs to involve general-level causal knowledge about the potential cause’s causal powers and temporal factors, which have to be gained from multiple observations. This, however, leads to the problem that a suitable type-level causal variable or reference class has to be chosen over which the relevant parameters can be computed.

In the present thesis it has not been tested how reasoners select reference classes to estimate the relevant causal parameters, but this is an interesting question for future studies. Take, for ex-ample, the question whether Peter’s smoking has caused his lung cancer. Peter might not only be categorized as a singular instantiation of the type “smoker” but also as a singular instantiation of the type “smoker consuming more than 10 but less than 50 cigarettes per day” or “smoker consuming more than

10 but less than 50 cigarettes per day and having cardiovascular problems”. The fact that singular cases can be subsumed under more or less abstract types raises the question which reference class people will tend to associate with a singular case when generating a singular causal judgment. An important problem here is that causal strength estimates in most cases will vary depending on the chosen reference class (Cartwright, 1989). A plausible hypothesis worth investigating is that people will prefer to associate singular cases with a reference class that is maximally homogeneous. This hypothesis is a plausible first pass but surely it needs empirical qualification. For example, narrowing the reference class too much (i.e., being too specific) so that the target case is its only member, would render the reference class useless. Another problem is that there is always a theoretically infinite number of reference classes if no further constraints on the features determining the class are introduced. We might additionally consider that Peter is a smoker who is around 40 years old, which seems like relevant information regarding his disease status, but we might also take into consideration that he has dark hair and large feet, which seems irrelevant. Thus, a more plausible hypothesis is that people choose the narrowest possible reference class that contains more than one individual and is described by features that are conceived as homogeneous with respect to all factors that are *generally* causally relevant for the target effect (see Cartwright, 1989; Waldmann & Hagmayer, 2001, for an elaboration on this problem).

7.2.4 Assessing the relevant parameters cross-sectionally vs. longitudinally

Another interesting question that follows from the considerations above and which should be addressed in future studies is whether reasoners prefer to assess the relevant causal parameters from cross-sectional (different entities observed at one point in time) or longitudinal observations (the same entity observed over time). Thus far, the Power Model of Causal Attribution is blind for whether the parameters (e.g., the causal power of target and alternative cause) are derived longitudinally or cross-sectionally, but a plausible hypothesis is that reasoners are sensitive to the way the parameters were obtained when they want to make singular causation judgments. As an example, imagine a person who one morning recognizes a slight muscle tremor in her hand after she had consumed a particular amount of coffee for breakfast, and that she now wondered whether it was the coffee that was responsible for the observed muscle tremor. In this situation, it seems plausible that she, to obtain an answer to her singular causation query, would not rely only on knowledge about the causal strength (and causal latency) between caffeine intake and muscle tremors

that she has gained from past observations of other people, but that she would also consider, and maybe even prefer, the contingency between caffeine consumption and muscle tremor that she gleaned from past observations of her own body behavior. A preference for assessing the causal parameters longitudinally will often be normative because moderator variables that introduce variability are more likely to remain constant in a longitudinal design.

7.3 Conclusion

Causal queries about singular cases arise frequently in our mundane lives and they are important in many professional disciplines such as the law, medicine, or engineering. Yet, the question of how it can be determined whether two events were actually causally connected turns out to be difficult to answer. The new Power Model of Causal Attribution that I proposed in this thesis combines information about causal power with temporal information about causal latency and onset differences, and incorporates uncertainty about the general causal relationship, in order to assess the probability with which an observed effect e was caused by a target cause c . The experiments showed that the new model better accounts for peoples' singular causation judgments than Cheng and Novick's (2005) original model, and thus provide evidence that the new model identifies important factors that had been missing. Yet, the theoretical considerations raised above also make clear that, to obtain a comprehensive account of singular causation judgments, different and important questions remain to be answered in the future.

References

- Ahn, W.-k., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.
- Allan, L. G., & Jenkins, H. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*, 381–405.
- Anderson, J. R. (2013). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Bes, B., Sloman, S., Lucas, C. G., & Raufaste, É. (2012). Non-bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*, 1178–1203.
- Bird, A. (2005). Abductive knowledge and holmesian inference. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (pp. 1–31). Oxford: Oxford University Press.
- Bird, A. (2007). Inference to the only explanation. *Philosophy and Phenomenological Research*, *74*, 424–432.
- Bird, A. (2010). Eliminative abduction: examples from medicine. *Studies in History and Philosophy of Science Part A*, *41*, 345–352.
- Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). The role of time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *44*, 1880–1910.
- Buehner, M. J. (2017). Space, time, and causality. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 549 – 564). New York: Oxford University Press.
- Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1119–1140.
- Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology Section A*, *56*,

- 865–890.
- Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*, 353–378.
- Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Clarendon Press.
- Cartwright, N. (2017). Single case causes: What is evidence and why. In H. K. Chao & J. Reiss (Eds.), *Philosophy of science in practice*. (pp. 11–24). Cham: Springer.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.
- Cheng, P. W., & Buehner, M. J. (2012). Causal learning. In K. J. Holyoak (Ed.), *The Oxford handbook of thinking and reasoning* (pp. 210–233). New York: Oxford University Press.
- Cheng, P. W., Liljehom, M., & Sandhofer, C. (2013). Logical consistency and objectivity in causal learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2034–2039). Austin, TX: Cognitive Science Society.
- Cheng, P. W., & Lu, H. (2017). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing the invariance of causal power. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 65–84). New York: Oxford University Press.
- Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.
- Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*, 694–706.
- Cheng, P. W., Novick, L. R., Liljeholm, M., & Ford, C. (2007). Explaining four psychological asymmetries in causal reasoning: Implications of causal assumptions for coherence. *Topics in contemporary philosophy*, *4*, 1–32.
- Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 201–215). New York: Oxford University Press.
- Dowe, P. (1995). Causality and conserved quantities: A reply to Salmon. *Philosophy of Science*, *62*, 321–333.
- Dowe, P. (2000). *Physical causation*. Cambridge: Cambridge University Press.
- Eells, E. (1991). *Probabilistic causality* (Vol. 1). Cambridge: Cambridge University

- Press.
- Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, *14*, 219–250.
- Glennan, S. (2017). *The new mechanical philosophy*. New York: Oxford University Press.
- Glymour, C. (2003). Learning, prediction and causal bayes nets. *Trends in Cognitive Science*, *7*, 43–48.
- Gopnik, A. (2000). Explanation as orgasm and the drive for causal understanding: The evolution, function and phenomenology of the theory formation system. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 299 – 323). Cambridge, MA: The MIT Press.
- Griffiths, T. L. (2017). Formalizing prior knowledge in causal induction. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 115 – 126). New York: Oxford University Press.
- Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). New York: Cambridge University Press.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.
- Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, *30*, 1128–1137.
- Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 225–276). Cambridge: The MIT Press.
- Halpern, J. Y. (2016). *Actual causality*. Cambridge, MA: MIT Press.
- Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*, 413–457.
- Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*, 843–887.
- Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. New York: Oxford University Press.
- Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*, 273–299.
- Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, *116*, 495–532.

- Hitchcock, C. (2009). Causal modelling. In H. Beebe, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 299 – 314). New York: Oxford University Press.
- Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*, 587–612.
- 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*, 702–727.
- Hume, D. (1748/1977). *An enquiry concerning human understanding*. Indianapolis: Hackett Publishing Company.
- Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.
- Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*, 1–17.
- Johnson, S. G. B., & Ahn, W.-k. (2017). Causal mechanisms. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 127–146). New York: Oxford University Press.
- Johnson, S. G. B., & Keil, F. C. (2018). Statistical and mechanistic information in evaluating causal claims. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 618–623). Austin, TX: Cognitive Science Society.
- Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*, 136 – 153.
- Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.
- Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 565–602). New York: Oxford University Press.
- Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856–876.
- Lagnado, D. A., & Sloman, S. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 451 – 460.
- Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*, 184–195.
- Lewis, D. (1973). Causation. *The journal of philosophy*, *17*, 556–567.
- Lewis, D. (1986). *Philosophical papers* (Vol. II). Oxford: Oxford University Press.
- Liljeholm, M., & Cheng, P. W. (2007). When is a cause the “same”? coherent

- generalization across contexts. *Psychological Science*, *18*, 1014–1021.
- Liljeholm, M., & Cheng, P. W. (2009). The influence of virtual sample size on confidence and causal-strength judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 157–172.
- Lipton, P. (2004). *Inference to the best explanation*. London: Routledge.
- Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? critique of cheng (1997). *Psychological Review*, *107*, 195.
- Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332.
- Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 415–432). New York: Oxford University Press.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–982.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–25.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. New York: Cambridge University Press.
- Mackie, J. L. (1974). *The cement of the universe: A study of causation*. New York: Oxford University Press.
- Marr, D. (1982). *A computational investigation into the human representation and processing of visual information*. San Diego, CA: Freeman.
- Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 433–457). New York: Oxford University Press.
- Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*, 277.
- Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1691 – 1696). Austin, TX: Cognitive Science Society.
- Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 806 – 812). Austin, TX: Cognitive Science Society.
- Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence.

- Psychological Review*, 111, 455—485.
- Paul, L. A. (2009). Counterfactual theories. In H. Beebe, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 158 – 184). New York: Oxford University Press.
- Paul, L. A., & Hall, E. J. (2013). *Causation: A user's guide*. New York: Oxford University Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, NY: Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. New York: Basic Books.
- Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, 145, 30–42.
- Reichenbach, H. (1978). The principle of causality and the possibility of its empirical confirmation. In *Hans Reichenbach selected writings 1909–1953* (pp. 345–371). Springer.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii: Current theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.
- Russo, F., & Williamson, J. (2011). Generic versus single-case causality: The case of autopsy. *European Journal for Philosophy of Science*, 1, 47–69.
- Salmon, W. C. (1965). The status of prior probabilities in statistical explanation. *Philosophy of Science*, 32, 137–146.
- Salmon, W. C. (1980). Probabilistic causality. *Pacific Philosophical Quarterly*, 61, 50-74.
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
- Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, 156, 164–176.
- Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, 41, 139–159.
- Shortle, J. F., Thompson, J. M., Gross, D., & Harris, C. M. (2018). *Fundamentals of queueing theory*. John Wiley & Sons.

- Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.
- Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2018). Assessing singular causation: The role of causal latencies. In T. Rogers, M. Rau, X. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1080–1085). Austin, TX: Cognitive Science Society.
- Stephan, S., Mayrhofer, R., & Waldmann, M. R. (submitted). The role of causal strengths and temporal information in singular causation judgments. *Manuscript submitted for publication*.
- Stephan, S., Tentori, K., Pighin, S., & Waldmann, M. R. (in preparation). Interpolating causal mechanisms: The paradox of knowing more.
- Stephan, S., & Waldmann, M. R. (2016). Answering causal queries about singular cases. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2795–2801). Austin, TX: Cognitive Science Society.
- Stephan, S., & Waldmann, M. R. (2017). Preemption in singular causation judgments: A computational model. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1126–1131). Austin, TX: Cognitive Science Society.
- Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*, 242–257.
- Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.
- Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland Publishing Company.
- Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.
- Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*, 27–58.
- Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 733–752). New York: Oxford University Press.
- Waldmann, M. R., & Mayrhofer, R. (2016). Hybrid causal representations. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 85–127). New York: Academic Press.
- White, P. (2017). Visual impressions of causality. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 245–264). New York: Oxford

University Press.

Appendix A

Summaries of the Additional Articles

The two articles that belong to the main project of this dissertation are included in Appendix B and Appendix C. I will here provide brief summaries of the additional papers which I wrote (or co-authored) during my PhD. The three proceedings articles in Appendices D, E, and F are direct precursors of the two main articles, and thus substantially overlap with the main articles. In Stephan and Waldmann (2016) (Appendix D), we proposed to embed Cheng and Novick’s (2005) Power Model of Causal Attribution into the Structure Induction Framework (Meder et al., 2014), in order to account for the problem of inferential uncertainty about the general causal structure and the parameters. The generalized attribution equation that was presented in Chapter 4 was for the first time proposed in Stephan and Waldmann (2017) (Appendix E). The formalization of the α parameter and the relevant temporal factors for its assessment were introduced in Stephan et al. (2018) (Appendix F).

A.1 Summary of Stephan, Tentori, Pighin, and Waldmann (in prep)

In a project (see Appendix G) in which Michael Waldmann and I collaborated with Katya Tentori and Stefania Pighin from the University of Trento, we investigated people’s causal belief updating in the context of causal interpolations. Causal interpolation describes the epistemic process in which a reasoner learns about the mechanism variables (M) linking two factors that have already been known to be causally related (thereby turning the representation of a direct relation, $C \rightarrow E$, into the representation of an indirect one, $C \rightarrow M \rightarrow E$). We were interested in how causal interpolation changes people’s predictive inferences. We found that causal interpolation leads to a systematic weakening in predictive probability judgments,

$P(e|c)$. In our paradigm, subjects first learned about the statistical relationship between a mutation of a fictitious gene, JPH3, and a fictitious disease, Lipogastrosis, and then were asked to estimate how likely it is that a randomly selected individual who carries the mutation also suffers from the disease. Thereafter, subjects were informed about the causal structure that generated the observed data. Subjects either learned that there was a direct relationship ($C \rightarrow E$) or that the mutation leads to the disease via a particular mechanism ($C \rightarrow M \rightarrow E$). Subjects then were again asked to estimate the predictive probability of disease given mutation. We found across all five experiments that subjects considered the effect to be less predictable once they had learned about the causal mechanism linking C and E . This weakening effect appears paradoxical, because it seems to imply that the more we learn about a causal relationship, the less we can say. The theoretical explanation for this weakening effect that we explored was that people might confuse interpolation scenarios with lengthening scenarios (or other types of causal network augmentations) in which causal variables are added on the effect side of a causal chain. In lengthening scenarios, where a reasoner first represents $A \rightarrow B$ and then learns $A \rightarrow B \rightarrow C$, the predictive probability $P(C|A)$ should indeed be weaker than $P(B|A)$ (unless the links are deterministic).

A.2 Summary of Nagel and Stephan (2015, 2016)

Together with Jonas Nagel I worked on a project (Appendix H and I) in which we investigated the role of mechanism information in causal explanations. In particular, we were interested to find out under which conditions an intermediate variable M connecting a root cause C with an effect E in a causal chain ($C \rightarrow M \rightarrow E$) is considered to be an “adequate mechanism” that explains how C generates E and under which conditions the intermediate variable M seems to take on the role of an *alternative* explanation of E that screens C off from E . Formally, the question was what the factors are that determine whether a causal relationship is perceived as being *transitive* or *intransitive*. Our general hypothesis was that whether an intermediate variable is regarded as an “adequate” mechanism or as an “alternative explanation” depends on the overall “sensitivity” of the relationship. The sensitivity of a causal relationship describes the degree to which it is robust toward perturbations of boundary conditions. Explanatory causes, according to this hypothesis, are those causes that not only bring about their effects in the narrow context of actually observed circumstances, but continue to do so in contexts in which different boundary conditions hold. One reason for why a causal relationship can be highly

sensitive (leading to intransitivity) is that the *mechanism* works reliably only under quite specific conditions, but is easily disturbed in other, similar situations. For example, in Experiment 1 in Nagel and Stephan (2015), subjects first learned about a strong statistical relationship between pupils' gender and their grades in a physical education class (e.g., all girls had good grades and all boys had bad grades). Subjects indicated that the pupils' gender was a good explanation for their grades in that class. Subjects then either learned that the relationship was mediated by a teacher who prefers girls over boys, or by physiological factors within the pupils (e.g., flexibility of the joints). We found that subjects in the teacher but not in the physiology condition revised their initial judgments concerning gender as an explanation for the grades in *this* class. Subjects in the teacher condition considered the pupils' gender to be less explanatory, and instead indicated that it was the teacher's preference that explained the relationship. No revision, by contrast, was found in the physiology condition. The sensitivity hypothesis explains this finding. If the influence of gender on grades is mediated by a genetically determined physiological mechanism, it likely will continue to hold in future observations with different samples of pupils. The teacher mechanism, by contrast, implies that this relationship depends on the presence of highly peculiar boundary conditions (a preferential bias) which will rarely be met in other, similar situations. The results of subsequent experiments, in which we eliminated different confounding factors (e.g., intentional agency, moral abnormality), confirmed the sensitivity hypothesis.

A.3 Summary of Stephan, Willemsen, and Gerstenberg (2017)

Together with Pascale Willemsen from the University of Bochum and Tobias Gerstenberg from the University of Stanford I have worked on a project (Appendix J) in which we investigated causal omissions. People often cite omissions (i.e., events that failed to occur) as causes of outcomes. Consider, for example, the situation where someone claims that “the ball went into the goal because the defender did not block it”. Causation by omission has been discussed heavily among philosophers because it seems to pose a problem for transference-based accounts of causality. Transference accounts appear to be incapable of handling causation by omission because no quantity is transferred from the cause (a negative event) to the effect (a positive event). Counterfactual accounts, by contrast, seem to be able handle these cases. According to a counterfactual analysis, the defender's not blocking the ball was the cause of its going into the goal if it is true that ball would not have gone into goal if the defender

had blocked it. However, many philosophers have pointed out that counterfactual accounts must deal with two problems here. First of all, there is the problem of causal *selection*. It also seems to be true that the ball would not have gone into the goal if someone else (e.g., the goalie) had blocked it. Why, then, does the defender seem to be the cause? A second problem is called the *underspecification problem*. In order to counterfactually identify an omission as a cause, a negative event has to be replaced with a positive event. The problem is that there are a infinitely many ways in which a negative event can be replaced by positive event. How can this set of possibilities be constrained?

In this project, we investigated how reasoners solve the *underspecification problem* in simple physical environments (involving marbles going or not going through a gate). We developed and tested a counterfactual simulation model that computes the probability with which an observed outcome (e.g., a red marble that failed to go through a gate) would have been different if the omission (a grey marble having failed to hit the red marble) had been replaced by a positive event (the red marble hitting the grey marble). The set of counterfactual possibilities that the model considers when it computes this probability is constrained by the physical properties of the environment. In two experiments, we found that the model's predictions closely mapped subjects' causality ratings, which suggested that subjects engaged in similar counterfactual simulations to derive their judgments. In future studies, we plan to test the model in more complex scenarios.

Appendix B

Stephan and Waldmann (2018)

Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science, 10*, 242 – 257.



Topics in Cognitive Science 10 (2018) 242–257
Copyright © 2017 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print/1756-8765 online
DOI: 10.1111/tops.12309

This article is part of the topic “Best of Papers from the Cognitive Science Society Annual Conference,” Wayne D. Gray (Topic Editor). For a full listing of topic papers, see [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1756-8765/earlyview](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview)

Preemption in Singular Causation Judgments: A Computational Model

Simon Stephan, Michael R. Waldmann

Department of Psychology, University of Göttingen

Received 18 September 2017; accepted 10 June 2017

Abstract

Causal queries about singular cases are ubiquitous, yet the question of how we assess whether a particular outcome was actually caused by a specific potential cause turns out to be difficult to answer. Relying on the causal power framework (Cheng, 1997), Cheng and Novick (2005) proposed a model of causal attribution intended to help answer this question. We challenge this model, both conceptually and empirically. We argue that the central problem of this model is that it treats causal powers that are probabilistically sufficient to generate the effect on a particular occasion as actual causes of the effect, and thus neglects that sufficient causal powers can be *preempted* in their efficacy. Also, the model does not take into account that reasoners incorporate uncertainty about the underlying general causal structure and strength of causes when making causal inferences. We propose a new measure of causal attribution and embed it into the structure induction model of singular causation (SISC; Stephan & Waldmann, 2016). Two experiments support the model.

Keywords: Singular causation; Causal attribution; Preemption; Causal reasoning; Bayesian modeling; Computational modeling

1. Introduction

Most people hold the belief that smoking causes lung cancer. Now, imagine that you learn that Peter, a passionate smoker, has contracted lung cancer. How strongly would you be willing to say that it was Peter’s smoking that was causally responsible for his disease?

Correspondence should be sent to Simon Stephan, Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany. E-mail: simon.stephan@psych.uni-goettingen.de

This example illustrates a scenario in which we seek an answer to a causal query about a singular case. Queries about singular causation are prevalent in everyday life and professional contexts, such as law or medicine. How do people derive causal judgments about singular cases? Of course, the mere fact that two factors C and E are *generally* causally connected (e.g., smoking often causing lung cancer) does not necessarily imply that a *singular* or *token* co-occurrence of these events (e.g., Peter's smoking and his lung cancer) manifests a causal relationship—a singular co-occurrence might be a mere *coincidence*. On the other hand, as singular causal connections are not directly observable, to what else than our general causal knowledge could we turn to obtain answers?

We are going to present a model that builds on the idea, first formalized by Cheng and Novick (2005), that the notion of unobservable *causal powers* (Cheng, 1997, see also Cartwright, 1989) plays an essential role in singular causation judgments. Yet we will demonstrate that Cheng and Novick's (2005) power PC model of causal attribution (CN model) makes assumptions that are not always plausible.

Generally, the CN model is intended to provide a normative answer to the question of how likely it is, or, in a Bayesian sense, how strongly one can believe that an observed outcome was actually caused by a potential cause factor. For example, for cases like the one above about Peter in which a potential cause c and an effect e have been observed, the CN model delivers the probability $p(c \rightarrow_{elc} e)$ with the arrow denoting a causal relation. We argue that the key conceptual problem of the model is that it treats target cause factors as singular causes whenever their causal powers are probabilistically sufficient to bring about the effect in a specific situation. While this appears to be at first sight a reasonable assumption, it ignores that the exact points in time at which different causes exert their powers play an important role in singular causation judgments (see Danks, 2017, for an overview): Crucially, causal powers that are sufficient to produce the effect on a particular occasion can be *preempted* by others, and in such cases they should not be held causally responsible for the occurrence of the outcome (e.g., Halpern & Pearl, 2005; Hitchcock, 2007). We will argue that preemption of target causes by background factors frequently occurs in singular causation scenarios and therefore presents a problem for the CN framework of causal attribution.

The neglect of the possibility of preemption through background causes represents a conceptual problem of the CN model. Another, rather epistemic, problem of the CN model is that it only uses point estimates to compute $p(c \rightarrow_{elc} e)$, and thus it does not take into account uncertainty about both the underlying general causal structure and the causal parameters (e.g., the size of the causal powers) that generate the observed data. To incorporate uncertainty about the causal parameters, Holyoak, Lee, and Lu (2010) have proposed a Bayesian variant of the CN model that uses probability distributions over the parameters instead of point estimates. However, their model still neglects uncertainty about the underlying general causal structure. Both sources of uncertainty have been demonstrated to influence causal learning and reasoning (Griffiths & Tenenbaum, 2005; Meder, Mayrhofer, & Waldmann, 2014). For this reason, Stephan and Waldmann (2016) proposed the structure induction model of singular causation (SISC) that takes into account both types of uncertainty for the computation of $p(c \rightarrow_{elc} e)$. Although three

experiments (Stephan & Waldmann, 2016) showed that SISC better accounted for the results than the standard power PC model of causal attribution, one shortcoming of the initial version of SISC was that it used the CN conceptualization of causal attribution that we criticize in the present article.

We will start with a theoretical section in which we defend a new measure of causal attribution as a component of SISC that is sensitive to the possibility of preemption through alternative causes. We then present the results of two experiments. Experiment 1a confirmed that singular causation judgments deviate systematically from the predictions of the CN model in line with our revised causal attribution equation. Experiment 1b assessed participants' notion of preemption using the same scenario as Experiment 1a. Importantly, Experiments 1a and 1b were designed such that uncertainty about the data generating general causal structure can be ruled out as an explanation for the observed results. Experiment 2 pursued two goals. First, we wanted to expose our revised SISC to a larger set of contingencies and compare its predictions with those made by further potential alternative models. Second, we introduced uncertainty about the general causal structure that produced the observed data to show that this factor does indeed influence singular causation judgments. The results of Experiment 2 showed that both a revision of the causal attribution equation and the consideration of statistical uncertainty are crucial to explain the findings.

2. The power PC model of causal attribution

According to Cheng's (1997) power PC theory (short for the causal-power theory of the probabilistic contrast model), causal power (or causal strength) is a hypothetical, unobservable entity that represents the strength with which causes bring about their effects. Mathematically, causal power is (in the generative case) expressed as the probability with which a target cause generates its effect in a hypothetical world in which all alternative observed and unobserved causes of the effect are absent. Because of the possibility of unobserved alternative causes, which can never be ruled out in the real world, causal power cannot be assessed directly and must therefore be inferred based on the observed covariation between the cause and effect factor and a set of background assumptions. Assuming that the target cause and background causes exert their powers independently of each other and that causal powers are independent of the base rates of the cause factors, the generative causal power w_c of a target cause C can be estimated with the following equation:

$$w_c = \frac{p(e|c) - p(e|\neg c)}{1 - p(e|\neg c)} = \frac{\Delta p}{1 - w_a}, \quad (1)$$

in which w_a represents the aggregate causal power of all alternative causes A of the effect, which is given by $p(e|\neg c)$. Δp is the common abbreviation for the difference term in the numerator of Eq. 1.

Under the causal Bayes net framework (Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 1993), the causal power w_c of a target cause factor C corresponds to the probabilistic weight w of the causal arrow that connects C with its effect E in a common effect structure in which C and the sum of all alternative causes, summarized in a single node A , combine in a *noisy-OR* gate (see S_1 in Fig. 2). Likewise, w_a corresponds to the weight of node A .

2.1. The CN measure of causal attribution

Cheng and Novick (2005) proposed several measures of causal attribution that apply to different cases. The measure of causal attribution for cases in which both c and e are present, as in the example above about Peter, utilizes the concept of causal power in the following way to deliver the conditional probability $p(c \rightarrow elc, e)$:

$$p(c \rightarrow e|c, e) = \frac{w_c}{p(e|c)} = \frac{w_c}{w_c + w_a - w_c \cdot w_a}. \quad (2)$$

Equation 2 shows that the CN model defines the probability with which c is causally responsible for e given that both have co-occurred by the fraction of the causal power of C and the conditional probability of the effect in the presence of C . Since the power PC theory assumes that C and A exert their causal powers independently of each other, $p(elc)$ can be rewritten as the sum of both causal powers minus their intersection (see second step of Eq. 2). Hence, what the CN model delivers is an estimation of the relative frequency of cases among all co-occurrences of C and E in which C 's causal power is probabilistically sufficient for the production of the effect. Our key criticism is that this relative frequency frequently overestimates the true proportion of cases in which we should actually causally attribute E 's occurrence to C because it neglects the possibility of preemption through background factors.

To illustrate the problem, consider the results of the fictitious experiment shown in Fig. 1 in which biologists investigated the influence of a chemical substance on the expression of a gene in laboratory mice. Because all mice in the test group that were treated with the substance ($p[elc] = 1$) but only half of the mice in the control group ($p[el-c] = .5$, the base rate) express the gene, the results provide strong evidence for the existence of a strong effect of the chemical. In fact, by applying Eq. 1 one can see that the causal power of the substance equals 1 in this case. Crucially, by applying Eq. 2, one can see that $p(c \rightarrow elc, e)$ also equals 1 in this case. What the CN model therefore prescribes is that we should attribute causal responsibility to the chemical whenever the chemical and gene expression are both present. In a Bayesian sense, the model suggests that a reasoner's degree of belief in a singular causal connection for an observed co-occurrence should be maximal.

But should we really be maximally confident that the expression of the gene in, for instance, Mouse # 25 must be causally attributed to the causal power of the chemical? If you have doubts, you were probably led by a prior assumption about the point in space and time at which the causal background factors A produced the observed base rate of 50%: In

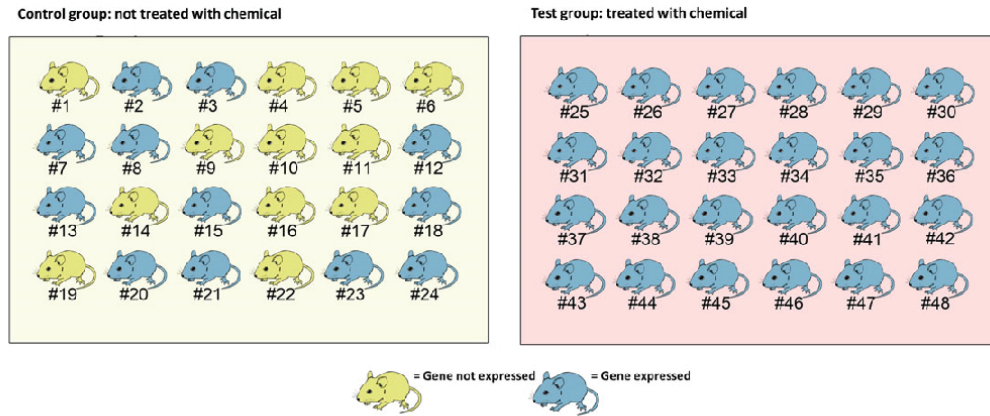


Fig. 1. Illustration of a hypothetical study testing the effect of a chemical on the expression of a gene. The control group is shown on the left and the test group treated with the chemical substance on the right. Mice having the gene expressed are depicted in blue.

the given scenario, it seems likely that these factors (e.g., transcription factors) already produced their effects *prior* to the introduction of the chemical. Under this assumption, however, it seems likely that not only 50% of the mice in the control group but also in the test group already expressed the gene prior to the study. Consequently, in those 50% of the mice it cannot be the chemical's causal power that is causally responsible for the effect because its causal efficacy has been preempted by the background factors.

3. A new measure of causal attribution

In the example given above it seems appropriate to say that the expression of the gene is caused by the chemical in only about half of the observed cases in which C and E have co-occurred. This conclusion is based on the assumption that in roughly half of the cases the causal power of the chemical has been preempted by the causal power of the background factors A . We propose a new measure of causal attribution that captures this intuition by refining Eq. 2 so that all cases among the joint occurrences of C and E for which the effect of C is assumed to be preempted by A are partialled out. This refined measure is given by

$$p(c \xrightarrow{\text{singular}} e|c, e) = \frac{w_c \cdot (1 - \alpha \cdot w_a)}{w_c + w_a - w_c \cdot w_a}, \quad (3)$$

in which we introduce α as a discounting parameter that represents the assumed probability with which the causal power of A preempts the causal power of C on occasions on which both are probabilistically sufficient to bring about the effect. For illustration, if we assume that A has caused the observed base rate of 0.5 prior to the application of the chemical, and if we assume further that A 's causal power has produced roughly equal

proportions of the effect in both groups, α takes on a value of 1. In this case, the point estimate for $p(c \xrightarrow{\text{sing.}} e|c, e)$ is about 0.5 instead of 1.0 that is predicted by Eq. 2. The difference between the two measures is that under the CN model the presence of two sufficient causal powers (w_c and w_a) is invariably conceptualized as a case of “symmetric overdetermination,” that is, a case in which the causal powers act exactly simultaneously, whereas the possibility that sufficient causes can preempt each other is neglected. Eq. 3 takes into account the possibility of preemption and thus delivers an estimation of the relative frequency of singular cases in which the target cause has *actually* been successful in generating the effect. In our view, preemption of C by a previously present background factor A seems to be prevalent in the cover stories typically reported in the literature (see, e.g., Griffiths & Tenenbaum, 2005). However, the discounting parameter α can be set to capture other cases. For example, cases of symmetric overdetermination or cases in which A is preempted by C could be modeled by setting α to 0. In these cases, our model and the CN model make identical predictions because Eq. 3 reduces to Eq. 2.

4. SISC: The structure induction model of singular causation

Apart from the fact that the CN model does not take into account the possibility that preempted causal powers should not be classified as singular causes, a further problem of the CN model is that it is insensitive to statistical uncertainty about both the underlying general causal structure and the size of the causal parameters. SISC¹ (Stephan & Waldmann, 2016) is sensitive to both types of uncertainty.

SISC was developed in the framework of causal Bayesian inference models (see also Griffiths & Tenenbaum, 2005; Meder et al., 2014); it takes observed data as evidence to update prior probabilities of mutually exclusive hypotheses. Under SISC, these competing hypotheses represent two causal structures that can account for a particular observed pattern of covariation. The two causal structures, S_0 and S_1 , are depicted graphically in Fig. 2. While there exists a causal arrow from C to E in S_1 , which indicates that C is a general cause of E , there is no causal arrow between C and E in S_0 . Furthermore, both models assume the existence of alternative causes of the effect, which are summarized in the single node A .

The core principle of SISC can be illustrated with Fig. 1. Assume someone suggests that S_0 is the causal structure that underlies the obtained results. Under this hypothesis, all observed co-occurrences of C and E would be mere coincidences. Yet, since the observed distribution of the events appears very unlikely to be due to mere sampling variation, S_0 is weakened as an explanation, while the alternative hypothesis, S_1 , is proportionally strengthened. In fact, the probability computed by SISC for S_1 for the data shown in Fig. 1 (i.e., the posterior probability of S_1 , $p[S_1|D]$) is almost 1. Now, imagine the same study had been conducted with a sample of merely eight mice but that $p(elt)$ and $p(elt-c)$ remain the same. In this case, it seems less certain that S_1 underlies these results. Smaller samples not only increase uncertainty about the underlying causal structure, they also impede the reliable estimation of the size of parameters (see Meder et al., 2014). SISC is sensitive to both types of uncertainty when estimating $p(c \xrightarrow{\text{sing.}} e|c, e)$.

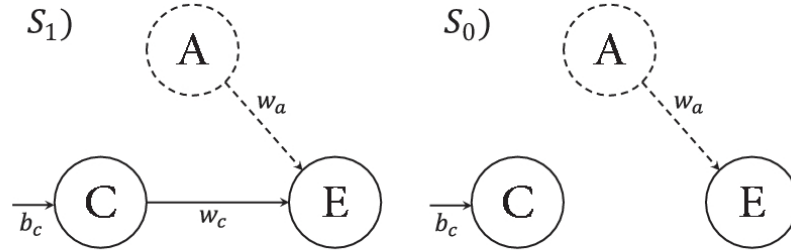


Fig. 2. The two causal structures considered by the structure induction model of singular causation as mutually exclusive explanations for observed patterns of covariation. C is a general cause of E under S_1 , whereas all co-occurrences of C and E are coincidental under S_0 . The parameters w_c and w_a denote the causal powers of the observed cause C and the unobserved background causes A , respectively; b_c denotes the base rate of C .

SISC implements different steps. First, it derives the posterior probabilities for each causal structure illustrated in Fig. 2. Applying Bayes' rule, the posterior probability for a causal structure is proportional to the likelihood of the data given the causal structure, weighted by the structure's prior probability:

$$p(S_i|D) \propto p(D|S_i) \cdot p(S_i). \quad (4)$$

$p(D|S_i)$ is the likelihood of the data given a particular structure, which is the integral over the likelihood function of the parameter values under the particular structure. $p(S_i)$ represents a structure's prior probability. Following Griffiths and Tenenbaum (2005) and Meder et al. (2014), the model initially assumes that both structures are equally likely, that is, $p(S_i) = 1/2$. When data become available, the posterior probabilities for each causal structure vary systematically with the observed contingency: the higher the contingency, the more likely S_1 becomes. Likewise, the posterior probabilities of the causal structures vary with the observed sample size. For example, given a fixed positive correlation, the posterior probability of S_1 increases the more data are observed.

Next, the model estimates the parameters b_c , w_c , and w_a , for each causal structure. To express parameter uncertainty, distributions rather than point estimates are inferred. The posterior probability distributions for the parameters, $p(w|D)$, are proportional to the likelihood of the data given the set of parameters w , weighted by the prior probability distributions of the parameters:

$$p(w|D) \propto p(D|w) \cdot p(w). \quad (5)$$

$p(D|w)$ is the likelihood of the data given the parameter values for b_c , w_c , and w_a . $p(w)$ is the prior joint probability of the parameters. The prior distributions of the parameters are independently set to flat, uninformative $beta(1,1)$ distributions. Since C does not cause E under S_0 , w_c is held fixed at 0 for this causal structure.

In the last step, SISC computes $p(c \xrightarrow{\text{sing.}} e|c, e)$ for each parameterized structure. The new discounting parameter alpha is set based on background assumptions about the target

scenario. For the scenarios we used in the present experiments, it is set to 1 because preemption through background causes seems to be highly probable. As all co-occurrences of c and e are coincidences under S_0 , $p(c \xrightarrow{\text{sing.}} e|c, e)$ is set to 0 for S_0 . For S_1 , Eq. 3 is applied. The final output of SISC is a single estimate for $p(c \xrightarrow{\text{sing.}} e|c, e)$, which is obtained through integrating out the two causal structures by summing over the derived values of $p(c \xrightarrow{\text{sing.}} e|c, e)$ for each structure weighted by its posterior probability:

$$p(c \xrightarrow{\text{sing.}} e|c, e; D) = \sum_i p(c \xrightarrow{\text{sing.}} e|c, e; S_i) \cdot p(S_i|D). \quad (6)$$

5. Experiment 1a

The goal of Experiment 1a was to test the revised measure of causal attribution used by SISC against the CN model of causal attribution for data sets with a sufficient cause, that is, $p(elt) = w_c = 1$, but varying base rates of the effect. Although the CN model predicts maximal confidence in singular causation assessments for any observed co-occurrence of C and E in this case, SISC predicts an influence of the base rate $p(e|¬c)$ under the assumption that A 's causal power generally preempts the effect of C . The goal of Experiment 1a was to demonstrate preemption as captured by our refined Eq. 3. To rule out uncertainty about the general causal structure as an explanation, we used sample sizes in our data sets for which the posterior probabilities of S_1 computed by SISC are close to 1. The predictions of the models are shown in Fig. 3. We set α in Eq. 3 to 1, which represents complete preemption of C by A . We also considered a Bayesian variant of the CN model that has been proposed by Holyoak et al. (2010). This model is sensitive to parameter uncertainty; it uses probability distributions over the parameters instead of point estimates. As Fig. 3 shows, the predictions of both variants of the CN model converge for large sample sizes because the influence of parameter uncertainty becomes minimal.

5.1. Methods

5.1.1. Participants

Ninety participants, all native English speakers, were recruited via the online platform Prolific (www.prolific.ac). Eighty-three participants ($M_{\text{age}} = 33.88$, $SD_{\text{age}} = 12.79$, 34 female) finished the experiment and were included in the statistical analyses. Participants received a monetary compensation of £0.60.

5.1.2. Design, materials, and procedure

Three contingencies (see Fig. 3) were manipulated between subjects with each participant responding to two causal test queries (general causation vs. singular causation). The general causation query was a general causal *structure* query referring to the assumed probability of the existence of a general effect of the cause factor. We included the

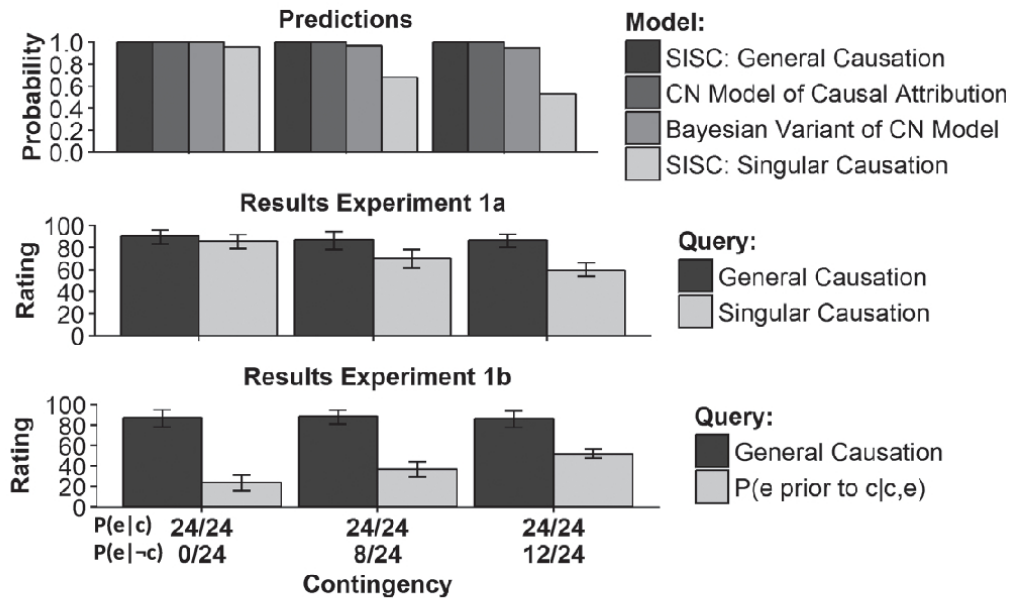


Fig. 3. Model predictions and results of Experiments 1a and b. The SISC predictions for general causation represent the posterior probabilities of S_1 . The CN model predictions for singular causation are obtained from Eq. 2. The Bayesian CN model uses probability distributions instead of point estimates. For the SISC predictions for singular causation, α in Eq. 3 was set to 1. The results show mean ratings and 95% bootstrapped CIs. Dark bars show general causation judgments; light bars show singular judgments.

general causation query to establish that uncertainty cannot account for the predicted pattern of singular causation ratings. The task was a standard elemental causal induction task. As cover story, we used the gene expression scenario (cf. Griffiths & Tenenbaum, 2005) mentioned above: Subjects were asked to assume that they were biologists who are interested in whether a particular chemical causes the expression of a particular gene in mice. Subjects read that they will be asked to conduct an experiment on the computer screen in which they will treat a random sample of mice with the substance, while a control sample will remain untreated. It was mentioned that the control sample is important as some mice may express the gene for other reasons.

Participants then were presented with an interactive animation showing the two samples arranged as in Fig. 1, and a pipette containing a reddish chemical substance. All mice had gray color in this animation. Participants then dragged and dropped the substance into the test group area, whereupon the background color changed to a light red to indicate successful application. On the next screen, subjects checked the results of the experimental manipulation by dragging a small magnifying glass over the mice in both groups. Mice that expressed the gene then became blue and those that did not express the gene became yellow. The final state of the animation looked like the illustration shown in Fig. 1.

Subsequently, participants responded to the two causal test questions. For the general causal structure query, participants were asked to indicate on a slider how confident they

were that the chemical has an effect on the expression of the gene (from “very certain that the chemical has no effect” to “very certain that the chemical has an effect”). The singular causation query asked subjects about Mouse #25 from the test group that had the gene expressed. Participants were asked to indicate on a slider how confident they were that it was the chemical substance that caused the expression of the gene in this single case (from “very certain that it was not the chemical” to “very certain that it was the chemical”).

5.2. Results and discussion

Fig. 3 shows the results.² As predicted by the posterior probability of S_1 , the general causation ratings were high ($M = 90.58$, $SD = 17.01$, $M = 87.55$, $SD = 22.12$, $M = 86.89$, $SD = 17.39$) across the three contingency conditions, indicating very little uncertainty about the underlying general causal structure. The singular causation ratings, by contrast, decreased with an increasing base rate of the effect ($M = 86.03$, $SD = 17.21$, $M = 70.41$, $SD = 22.28$, $M = 59.61$, $SD = 16.96$), as predicted by SISC but not by the two CN models. A multilevel model analysis revealed significant main effects for type of causal query, $\chi^2(1) = 32.45$, $p < .001$, contingency, $\chi^2(1) = 12.63$, $p < .01$, and the interaction of query \times contingency, $\chi^2(1) = 13.10$, $p < .01$. Fig. 3 shows that the main effect of causal query was obtained because general causation ratings were, on average, higher than singular causation ratings. Furthermore, Fig. 3 shows that the main effect of contingency is driven by the selective decrease observed for the singular causation ratings and the fact that there were only small differences observed for the general causation ratings. As predicted by SISC, a quantitative trend analysis computed for the general causation ratings yielded no significant trends, $F(1,80) = 0.50$. For the singular causation ratings, by contrast, a quantitative trend analysis yielded a significant negative linear trend, $F(1, 80) = 25.93$, $p < .001$, $r_{\text{effect size}} = .49$. In sum, both the trends for general as well as for singular causation ratings are captured well by SISC. The observed trend for the singular causation judgments is, however, neither predicted by the CN model using point estimates nor by the Bayesian extension incorporating parameter uncertainty.

6. Experiment 1b

The goal of Experiment 1b was to assess how likely participants think it is that a particular mouse from the test group already showed the effect due to background factors prior to the occurrence of the target cause. Thus, instead of singular causation judgments for a particular individual, we asked subjects to estimate $p(e \text{ prior to } clc, e)$. Crucially, responses to this query provide us with an estimate of the α value in Eq. 3 that participants assumed: the higher the assumed value of alpha, the closer should the ratings be to the observed base rate of the effect. This prediction can be derived from Eq. 3, in which w_a , which corresponds to the base rate, is weighted by α .

6.1. Methods

6.1.1. Participants

A total of 88 participants, all native English speakers, were recruited via Prolific (www.prolific.ac). Seventy-four participants ($M_{\text{age}} = 32.12$, $SD_{\text{age}} = 11.51$, 36 female) finished the experiment and were included in the statistical analyses. Participants received a monetary compensation of £0.60.

6.1.2. Design, materials, and procedure

The study design and the materials were the same as in Experiment 1a. The only difference was that, instead of a singular causation judgment for Mouse #25 which was treated with the chemical and expressed the gene, we asked participants how likely they think it is that this mouse had already expressed the gene prior to the experiment. To keep things as similar as possible to Experiment 1a, we also asked participants to answer a general causal structure query.

6.2. Results and discussion

We replicated the results for the general causation judgments found in Experiment 1a ($M = 87.26$, $SD = 21.70$, $M = 88.63$, $SD = 17.70$, $M = 86.37$, $SD = 22.88$, see Fig. 3). A quantitative trend analysis for the general causation judgments was not significant, $F(1, 71) = 0.02$. However, the probability judgments about the presence of the effect prior to the application of the chemical in the single case showed the opposite trend as the singular causation judgments in Experiment 1a ($M = 24.04$, $SD = 19.08$, $M = 37.00$, $SD = 19.19$, $M = 51.78$, $SD = 11.16$). This finding supports our hypothesis that assumptions about preemption influence singular causation judgments, as predicted by Eq. 3. A quantitative trend analysis yielded a significant positive linear trend, $F(1,71) = 34.40$, $p < .001$, $r_{\text{effect size}} = .57$. Furthermore, as the ratings approximately corresponded to the observed base rate of the effect (except for the zero-base rate condition in which they were higher), the results indicate that participants indeed assumed high α values.

7. Experiment 2

Experiment 1a showed that singular causation ratings for sufficient causes deviate systematically from the predictions of the CN models. This deviation is predicted as a consequence of assumptions about preemption relations between C and A . Experiment 2 pursued two main goals: first, we aimed to test SISC using a larger set of contingencies with a combination of different levels of $p(elt)$ and $p(elt-c)$. Second, we wanted to demonstrate that parameter and structure uncertainty not only influence general causation (Griffiths & Tenenbaum, 2005) but also singular causation judgments. We therefore used the set of contingencies studied in Buehner, Cheng, and Clifford (2003). We excluded the one contingency from the set in which the effect never occurs as it does not make sense

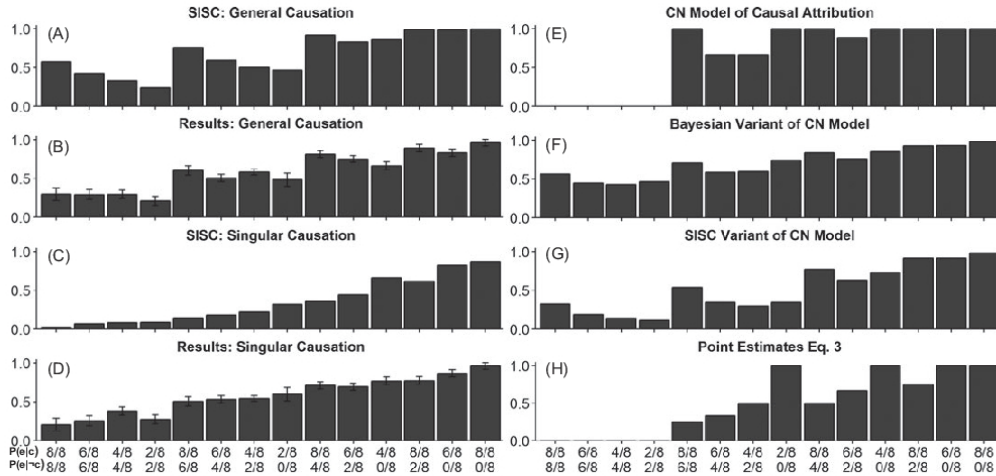


Fig. 4. Predictions of different models and results (means and within-subjects adjusted 95% CIs) of Experiment 2. Graphs (A) and (B) refer to general causation assessments. All other graphs refer to singular causation assessments. The SISC variant of the CN model incorporates structure and parameter uncertainty but applies Eq. 2. Graph (H) shows point estimates of our refined Eq. 3 and was included to underline that general causation uncertainty needs to be taken into account.

to ask for singular causation if the effect is absent. The observed samples of mice were smaller than in Experiment 1a, which should lead to increased uncertainty. The data sets and model predictions are shown in Fig. 4. For the singular causation predictions made by SISC, we set the discount parameter α to 1 again.

7.1. Methods

7.1.1. Participants

A total of 82 participants ($M_{\text{age}} = 34.41$, $SD_{\text{age}} = 10.42$, 31 female) participated in this online study and finished the experiment. They were paid £1.00 for their participation. As in Experiments 1 and 2, participants were recruited via Prolific (www.prolific.ac). All participants were native English speakers.

7.1.2. Design, materials, and procedure

The causal query (general causation vs. singular causation) was manipulated between subjects. The 14 different contingency data sets were varied within subject. They were presented in random order. We used the same cover story as in Experiment 1, except that subjects read that they will investigate the effects of 14 different chemicals on 14 different genes in 14 different samples of mice. We pointed out in the instructions that the results of the different studies are independent of each other. The instructions were similar to those used by Griffiths and Tenenbaum (2005) in their Experiment 1. The procedure was similar to Experiments 1a and 1b. As in Experiments 1a and 1b, the assignment

of mice to the cells of the contingency tables was randomly determined. In the present experiment, the test mouse for the singular causation query showing both c and e was randomly chosen prior to the experiment.

7.2. Results and discussion

Fig. 4 shows the results and the predictions of the different models: (A) and (B) display the predictions of SISC for general causation and the mean general causation responses. Panels (C) and (D) show the predictions of SISC and the results of the singular causation queries. Predictions of the standard CN model and its Bayesian variant are displayed in (E) and (F). Graph (G) shows predictions of SISC when α is set to zero but both structure and parameter uncertainty are incorporated. Finally, (H) shows point estimates of Eq. 3 while neglecting statistical uncertainty.

The overall pattern for both general and singular causation ratings was captured best by the revised version of SISC. As in our previous experiments, the results show that participants differentiated between general and singular queries. Moreover, the responses to the general causation query replicate those analyzed by Griffiths and Tenenbaum (2005). Most important, the singular causation assessments were captured best by the revised SISC shown in graph (C). The other models, by contrast, especially struggled to account for the local trends observed within the subsections of the contingency set in which Δp is constant. The difference between the singular causation ratings and the point estimates obtained from our revised equation for causal attribution (Eq. 3) implies that participants were, in addition to the possibility of preemption, sensitive to structure and parameter uncertainty.

A multilevel model analysis confirmed the main effect of contingency, $\chi^2(13) = 1010.04$, $p < .001$, as well as the interaction between contingency \times query, $\chi^2(13) = 49.39$, $p < .001$, that is shown in Fig. 4. To test the different models, we computed different fit measures shown in Table 1. As can be seen there, SISC achieved a good fit in the overall fit measures (bottom part of the table). It explained most variance, with $R^2 = .88$, and yielded the second smallest *RMSE* of .11. Yet all models obtained relatively high values on the global measures. Even the CN model with the lowest overall fit accounted for 77% of the variance. The similarity between the models is not unexpected, however, as all models are sensitive to Δp . More interesting are the fit measures for the subsections of the contingency set in which Δp is kept constant. The upper part of Table 1 shows that SISC yielded high fit values there, too, and hence accounted well for these local trends, whereas the Bayesian CN model, which yielded the smallest *RSME*, even yielded negative correlations.

8. General discussion

We addressed two different problems that the power PC framework of causal attribution (Cheng & Novick, 2005) faces: first, the CN model attributes causal responsibility

Table 1
Model comparisons for singular causation judgments in Experiment 2

Fit Measure	SISC	CN Model	Bayesian CN Model	SISC CN Model	Point Est. Eq. 3
$r_{\Delta p=.00}$.72	N/A	-.78	-.68	N/A
$r_{\Delta p=.25}$	1.00	.21	.38	-.61	.98
$r_{\Delta p=.50}$.88	.68	.79	.44	.85
$r_{\Delta p=.75}$	1.00	N/A	1.00	-1.00	1.00
$M_{r\Delta p}$.90	.44	.35	-.46	.94
$r_{overall}$.94	.88	.93	.90	.90
R^2	.88	.77	.87	.82	.82
$RMSE$.11	.27	.09	.14	.22

Notes. Δp refers to the different contingency levels (.00, .25, .75, 1.00) within the whole data set; $r_{\Delta p}$ expresses the model fits for these levels. N/A represents undefined values.

for the occurrence of a particular effect e to a present singular event c whenever its causal power is sufficient to bring about the effect on this occasion. We have argued that this conceptualization fails to take into account that people make assumptions about the point in time at which different causal powers exert their influences. Not every manifestation of a sufficient cause c needs to be causally responsible for an observed outcome; it might be the case that a competing cause (i.e., a) preempts it. This problem of *redundant* causation, which occurs whenever two causes are individually sufficient for the effect, is widely acknowledged in the philosophical literature as a challenge for counterfactual accounts of causation (see, e.g., Paul & Hall, 2013). To account for the possibility of preemption, we have modified the equation developed by Cheng and Novick (2005) as an account of causal attribution. The revised equation includes the discount parameter α that can be set to express domain-related assumptions about the temporal relations between the alternative causal factors. A second shortcoming of the standard causal attribution model (Cheng & Novick, 2005) is that it does not take into account statistical uncertainty about structure and causal parameters (cf. Griffiths & Tenenbaum, 2005; Meder et al., 2014). Our model SISC remedies both shortcomings. It is sensitive to both the temporal relations between the alternative causes and to statistical uncertainty. Our experiments showed that both aspects are important to account for subjects' judgments about singular causation.

We have set the discount parameter α to 1 in Eq. 3, which implies a preemption relation between A and C whenever A 's and C 's causal power are both sufficient in a situation. In future experiments, we plan to manipulate α by manipulating domain assumptions about the temporal relations between C and A . Cases in which α is 1 are situations in which A always preempts C . The cover stories used in the present experiments are an example in which it is plausible to assume that A represents a temporally stable factor that has already been efficacious prior to the manipulation of C . Although preemption through A seems to be the predominant scenario in the world, there certainly are cases in which other assumptions need to be made. Consider cases of symmetric over-determination that have also been discussed in the literature (see Paul & Hall, 2013): in the

famous firing squad scenario, for example, in which each shooter is treated as a sufficient cause for the death of the target, the intuition is that each shooter should be counted as a singular cause of the death of the victim. Our model can be applied to those cases, too. In this case, α would have to be set to zero. Similarly, α would have to be set to zero in scenarios in which reasoners have a high degree of belief that C preempts A so that A cannot manifest its potential causal power. Furthermore, in scenarios in which the preemption relation is highly uncertain, α might be set to one half. Cases of temporal variability between C and A might also be an interesting topic for future studies.

An important source of information that might aid singular causation judgments and that has been suggested in the literature (e.g., Cartwright, 2017) is causal mechanism information. For example, the observation of tar in the lung and genetic alterations might strengthen our confidence that Peter's smoking did indeed cause his lung cancer. Mechanism information can be modeled by inserting intermediate nodes in the path connecting C to E , thus forming a causal chain. In future experiments, we plan to use SISC to model such cases. The present studies are just a first step in the direction of providing a full account of singular causation judgments.

Acknowledgments

We thank Jonas Nagel and Ralf Mayrhofer for helpful discussions. This research was supported by Deutsche Forschungsgemeinschaft (DFG) Grant WA 621/24-1

Notes

1. Information on how to access the program code can be found in the Supplementary Materials.
2. The data of the reported experiments are provided in the Supplementary Materials.

References

- Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1119–1140.
- Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford, UK: Oxford University Press.
- Cartwright, N. (2017). Single case causes: What is evidence and why. In H.-K. Chao & J. Reiss (Eds.), *Philosophy of science in practice: Nancy Cartwright and the nature of scientific reasoning* (pp. 11–24). Cham, UK: Springer International Publishing.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104, 367–405.
- Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, 112, 694–706.

- Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 201–215). New York: Oxford University Press.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334–384.
- Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part i: Causes. *The British Journal for the Philosophy of Science, 56*, 843–887.
- Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review, 116*, 495–532.
- 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*, 702–727.
- Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*, 277–301.
- Paul, L. A., & Hall, E. J. (2013). *Causation: A user's guide*. Oxford, UK: Oxford University Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan-Kaufmann.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
- Spirtes, P., Glymour, C. N., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.
- Stephan, S., & Waldmann, M. R. (2016). Answering causal queries about singular cases. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 2795–2801). Austin, TX: Cognitive Science Society.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:
Data S1. SISC Code and Experimental Data

Appendix C

Stephan, Mayrhofer, and Waldmann (subm.)

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (submitted). The role of causal strength and temporal information in singular causation judgments. *Manuscript submitted for publication.*

The Role of Causal Strength and Temporal Information in Singular Causation
Judgments

Simon Stephan, Ralf Mayrhofer, and Michael R. Waldmann
University of Göttingen

Author Note

Simon Stephan, Ralf Mayrhofer, and Michael R. Waldmann, Department of Psychology, University of Göttingen.

Experiments 2 and 3 were presented at the 2018 Annual Conference of the Cognitive Science Society in Madison, Wisconsin.

This research was supported by Deutsche Forschungsgemeinschaft (DFG) Grants WA 621/24-1 and MA 6545/1-2, the latter as part of the priority program “New Frameworks of Rationality” (SPP 1516).

Correspondence concerning this article should be addressed to Simon Stephan, Department of Psychology, University of Göttingen, Gosslerstr. 34, 37073 Göttingen, Germany.

Email: simon.stephan@psych.uni-goettingen.de

Abstract

Causal queries about singular cases, which inquire whether actual individual events are causally connected, are prevalent in daily life and important in professional disciplines, such as the law, medicine, or engineering. Because causal links cannot be directly observed, singular causation judgments require an assessment of whether a co-occurrence of two events c and e is causal or simply coincidental. How can this decision be made? Relying on previous work by Cheng and Novick (2005) and Stephan and Waldmann (2018) we propose a model that combines information about the causal strengths of the potential causal events with information about their temporal relations to derive answers to singular causation queries. The relative causal strengths of the potential cause factors are relevant because weak causes are more likely to fail to generate effects than strong causes. But even a strong factor needs not necessarily be causal in a singular case, for it could have been preempted by an alternative cause. We here show how information about causal strength and about two different temporal components, the potential causes' onset times and their causal latencies, can be formalized and integrated to account for the possibility of causal preemption and to assess singular causation. Four experiments were conducted in which we tested the validity of the model. The results showed that people integrate the different types of information as predicted by the new model.

Keywords: singular causation; causal attribution; preemption; time; causal reasoning; computational modeling

1 Introduction

The ability to reason causally about how the world works is one of our central cognitive capacities (see Sloman, 2005; Waldmann, 2017). Whereas previous research has frequently investigated the question of how we learn and reason about causal events that refer to classes of objects or people (e.g., “smoking causes lung disease”), little is known about how we reason about causal explanations of singular events (but see Danks, 2017; Lombrozo & Vasilyeva, 2017). For example, if a smoker is diagnosed with lung cancer, we might wonder whether it has been her smoking that caused her lung cancer or whether the disease was due to something else, such as exposure to asbestos. In the context of law, we may be interested in whether a defendant actually caused the death of his fiancée when he was drunk, which is a different query from the one of a sociologist who is interested in the general relation between alcoholism and violent acts (see Hart & Honoré, 1959/1985; Lagnado & Gerstenberg, 2017). All these queries are aided by knowledge about general causal relations, but cannot be reduced to them. Knowing that smoking probabilistically causes lung cancer does not entail that the lung cancer of a specific smoker was actually caused by her smoking. It may very well be a coincidence.

More abstractly, when assessing singular causation reasoners need to take into consideration that a singular co-occurrence c and e of the general factors C and E , even if these are already known to be generally causally connected ($C \rightarrow E$), could well have been a mere coincidence. The present research asks how reasoners form beliefs about causal links in singular cases. If potential causes (say c or/and a) and a particular outcome (e) are known to have co-occurred, what influences the degree to which a reasoner believes that there has been a causal connection between them?

In the present paper we present a new computational model of singular causation judgments that extends previous work of Cheng and Novick (2005) and Stephan and Waldmann (2018). We will show that the assessment of singular causation relations requires the combination of two types of information: (a) information about the causal strengths of the potential cause factors, and (b) information about temporal relations

between causes and effect. We will focus on two types of temporal information that are important: One type is information about the onset times of the potential causes and the other is information about causal latencies, by which we mean the time it takes causes to bring about their effects. We will show how information about these components can be formalized and how the generalized model of causal attribution combines causal strength and temporal information to predict singular causation judgments. The results of four experiments will be reported which supported the predictions of the new model.

1.1 General causal relations

All models of singular causation that will be discussed here postulate a tight link between general and singular causation. To answer the question whether Peter's lung cancer is caused by his smoking, for example, it is necessary to know that generally smoking tends to cause lung cancer. Moreover, the stronger the causal relationship is, the more likely it is that the singular case is causal rather than a coincidence. A popular theory of how causal strength of a general relation can be estimated is Cheng's (1997) power PC theory. According to Cheng, causal power is the probability of an effect given a target cause in the hypothetical absence of alternative observed or unobserved causes. Since such cases cannot be directly observed, causal power needs to be estimated based on the observed covariation between cause and effect.

For an illustration of Cheng's (1997) method of estimating strength, consider the upper part of Fig. 1, which shows an example contingency that might have resulted from a medical study in which doctors investigated whether a newly developed drug (C) can cause headaches as a side effect (E) (cf. Liljeholm & Cheng, 2007). The right panel shows the results that were obtained in the control group (e.g., treated with a placebo), while the left panel shows the results that were obtained in the treatment group. As can be seen, 12 out of the 24 subjects in the control group suffered from headaches (at the time the researchers made the assessment), yielding $P(e|-c) = 0.50$. In the treatment group, by contrast, 18 out of the 24 subjects suffered from headaches,

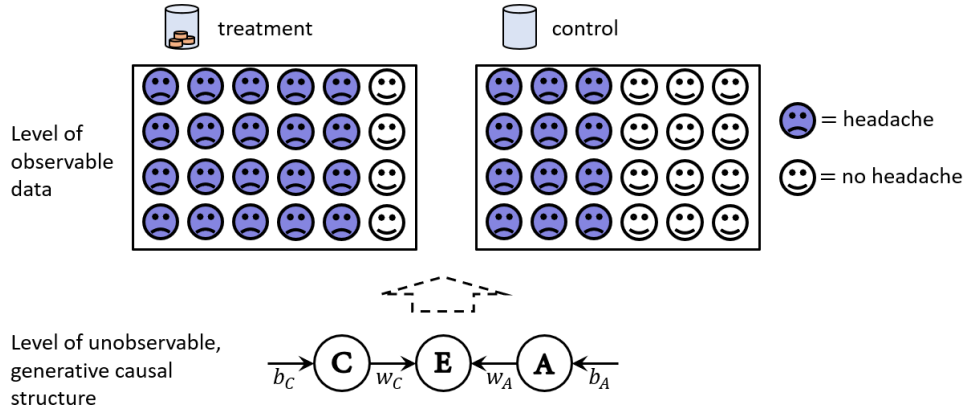


Figure 1. Relation between observable covariation data and the unobservable causal structure assumed to underlie the data. C and E in the causal structure denote the target cause and effect factor, respectively. A is assumed to comprise the sum of unobserved alternative causes of E . b_C and b_A denote the base rates of C and A , w_C and w_A the causal powers of C and A .

yielding $P(e|c) = 0.75$. The results show a positive probabilistic contrast of $\Delta P = 0.75 - 0.50 = 0.25$, indicating that the probability of headaches in the presence of the newly developed medicine increases by 25 percent.

ΔP measures the degree of covariation between two factors but the degree of covariation does not always reflect the causal strength of a factor. Cheng (1997) has shown under which circumstances the (generative or preventive) causal power (or strength) of a target cause factor C can be estimated from observable covariation data. A graphical representation of the unobservable causal structure that reasoners would assume to underlie the observable contingency ΔP between a target cause C and a target effect E is shown in the bottom part of Figure 1. The power PC theory principally distinguishes between two generative causal influences on E : the power exerted by the target cause C , and the power exerted by the sum of all (unobserved) alternative causes A of E . In the simplest case, A consists of only one alternative cause. The power PC theory assumes that C and A occur independently of each other with certain base rates, denoted b_C and b_A , respectively. Furthermore, it is assumed that C

and A exert their powers on E independently of each other (i.e., non-interactively). The functional form of the causal structure established by these assumptions is called a *noisy-OR gate* in the causal Bayes net literature (Glymour, 2003; Griffiths & Tenenbaum, 2005; Meder, Mayrhofer, & Waldmann, 2014; Pearl, 1988, 2000). The causal powers of C and A can be labeled w_C and w_A , respectively, and correspond to the probabilistic weights of the causal arrows connecting the causes with their effect (see Griffiths & Tenenbaum, 2005). Importantly, if the assumptions underlying power PC theory are met, the theory provides a causal explanation of the observed ΔP . According to the theory, ΔP is given by $\Delta P = w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A - b_A \cdot w_A$, where the first term, $w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A$, corresponds to $P(e|c)$ and the last term, $b_A \cdot w_A$, corresponds to $P(e|\neg c)$. By substituting $b_A \cdot w_A$ with $P(e|\neg c)$ and by re-arranging the equation, the causal power of a generative target cause C can be estimated with the following equation:

$$w_C = \frac{\Delta P}{1 - P(e|\neg c)}. \quad (1)$$

For an illustration, consider again the results of the fictitious study depicted in the upper part of Fig. 1. According to causal power theory, the 12 instances of headaches that were observed in the control group must have been caused by some unobserved background factor(s), with probability $b_A \cdot w_A = 0.50$. As the theory assumes that both the base rates and the causal powers of C and A are independent of each other, the 18 instances of headaches observed in the treatment group must, by contrast, be explained by the joint influence of C and A , whereby A 's contribution is already known to have been $b_A \cdot w_A = 0.50$. Hence, 12 of the 18 instances of headaches that occurred in the treatment group can be causally explained by the influence of A . Only the remaining 12 subjects who received the treatment C could actually reveal its causal power. 8 out of these 12 subjects suffer from headache. The causal power of the drug thus is $w_C = \frac{8}{12} = \frac{2}{3}$, implying that a patient not suffering from headaches who is given the newly developed medicine can expect to develop headaches with a 66 percent probability.

2 Cheng and Novick’s (2005) causal power model of attribution

The last section described how causal power can be estimated for general causal relations. Cheng and Novick (2005) have shown how general-level causal knowledge can be applied to answer queries about singular instantiations (see also Holyoak, Lee, & Lu, 2010). Cheng and Novick derived equations for different types of queries; we will focus here on singular causation queries in which c and e have been observed to be present.

Imagine a doctor who has diagnosed lung cancer in one of her patients knows that the patient is a smoker. According to Cheng and Novick (2005), what needs to be considered for estimating the probability that c (“the patient’s smoking”) actually caused the observed effect e (“her lung cancer”) is the probability with which C tends to cause E (i.e., its causal power) and the probability of E given C . This quantity can formally be denoted as $P(c \rightarrow e|c, e)$ and can be estimated with Eq. 2. This equation is supposed to filter out the cases of e in the presence of c that are actually caused by c using the causal power estimate w_C .

$$P(c \rightarrow e|c, e) = \frac{w_C}{w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A} = \frac{w_C}{P(e|c)}. \quad (2)$$

Furthermore, one can also imagine a situation involving exactly two potential causes C and A of a common effect E . Imagine it was known that both potential causes and the effect were present (c, a, e) and that a reasoner wanted to know which of them is more likely to be the singular cause of e . In this case, $P(c \rightarrow e|c, a, e)$ is provided by:

$$P(c \rightarrow e|c, a, e) = \frac{w_C}{w_C + w_A - w_C \cdot w_A} = \frac{w_C}{P(e|c, a)}. \quad (3)$$

Equations 2 and 3 describe how different situations in which causal queries about singular cases can arise should be handled according to the power PC framework of causal attribution. Fig. 2 illustrates the power PC theory of causal attribution graphically. For simplicity, imagine Fig. 2 depicted a scenario in which there exists one alternative cause factor A in addition to the target cause factor C . A is assumed to be present in both the treatment and the control group. Now imagine a patient showing the effect (e) who was randomly sampled from the treatment group (c, a, e). What

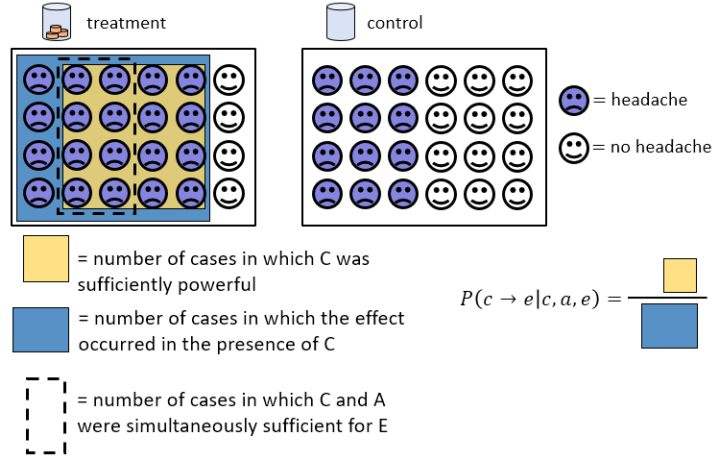


Figure 2. Illustration of the power PC model of causal attribution. The blue area represents the relative frequency of cases in which the E occurs when C and A are present. The yellow area represents the relative frequency of cases in which C 's causal power was probabilistically sufficient to generate E . Cases within the dashed frame represent those in which C and A were both sufficient.

would, according to the power PC framework of causal attribution, be the probability with which the effect e was actually caused by c , $P(c \rightarrow e|c, a, e)$? According to Eq. 3, $P(c \rightarrow e|c, a, e)$ can be estimated by normalizing w_C by the probability of E given C and A , $P(e|c, a)$. The number of effects in the presence of C and A is highlighted in Fig. 2 by the blue area in the treatment-group panel. As can be seen, the blue area marks those 20 out of the 24 patients who took the medicine and showed the effect, that is, $P(e|c, a) = \frac{20}{24} = \frac{5}{6}$. The causal power of C is highlighted by the yellow area in the treatment-group panel. From Eq. 1, we already know that the causal power of C is $w_C = \frac{2}{3}$. C was therefore probabilistically sufficient to produce E in 16 out of the 24 patients who took the medicine. $P(c \rightarrow e|c, a, e)$ can be understood as the ratio of the yellow and the blue area in the treatment group panel.

2.1 The insufficiency of probabilistic sufficiency: cases of causal preemption

Stephan and Waldmann (2018) have pointed out in a recent article that the causal power theory of causal attribution has one important deficit. The theory fails to

capture situations of “probabilistic overdetermination” in which a sufficient potential target cause factor was preempted in its efficacy by an alternative cause that was also present and sufficient on the occasion. For an illustration, consider again the treatment group shown in Fig. 2. Eq. 1 revealed that C was probabilistically sufficient in 16 out of the 24 treatment-group subjects (highlighted by the yellow area). But we also know from the control group that the alternative cause A tends to produce the effect with a causal power of $w_A = 0.50$. Under the independence assumptions of the power PC theory we can hence assume that A was also probabilistically sufficient in fifty percent of the cases in the treatment group, that is, in 12 of the 24 subjects. Hence there is a “sufficiency overlap” between C and A of $w_A \cdot w_C = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$. There are eight subjects in the treatment group in which C and A were simultaneously sufficient. Imagine that these eight overdetermined cases were the ones that are marked by the dashed rectangle shown in Fig. 2. Eq. 3 attributes all these cases to C . Thus, $P(c \rightarrow e|c, a, e)$ tends to overestimate the relative frequency of effects singularly caused by C , as for some cases the alternative cause a could have preempted the target cause c in generating the effect.

A standard example of preemption, often discussed in the philosophical literature (e.g., Halpern, 2016; Halpern & Hitchcock, 2015; Lewis, 1973; Pearl, 2000) is a scenario involving two perfectly accurate rock throwers, Billy and Suzy. It is assumed in this example that neither Billy nor Suzy, when acting alone, ever fail to hit and destroy a bottle. On a particular occasion, however, Billy and Suzy both end up aiming for the same bottle and both are throwing their rocks with nearly identical speed. Suzy, however, manages to throw her rock a bit earlier than Billy, and the bottle breaks before Billy’s stone reaches its target. In this case it is intuitively Suzy’s but not Billy’s throwing that caused the bottle’s breaking, no matter that Billy’s throw was also sufficiently accurate and powerful.

Singular cases involving preemption can be illustrated schematically with neuron diagrams as shown in Fig. 3 (see Paul & Hall, 2013), where a situation is depicted in which the potential cause c was preempted by an alternative cause a . Such situations of preemption are perfectly in line with the general causal structure assumed by the power

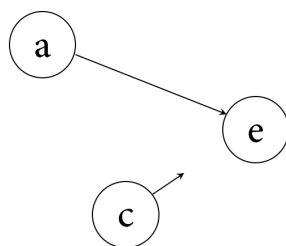


Figure 3. Neuron diagram modeling a situation of causal preemption (cf. Paul & Hall, 2013). The interrupted causal arrow departing from c and the arrow connecting a and e illustrate a situation in which c was causally preempted by a .

PC framework. Yet, as the power PC model of causal attribution considers only the relative sufficiency of the potential causes, it cannot distinguish between different situations of preemption. Applying the preemption scenario to the fictitious medical study, imagine that in Fig. 2 the background factor had produced all its effects before the subjects in the treatment group received the medicine. In this case, the eight occurrences of the effect identified by the dashed rectangle should intuitively not be attributed to the target cause but to the alternative cause. All singular causal relations that are compatible with the general causal structure and a singular observation in which the two potential causes and the effect were present (c, a, e) are depicted in Fig. 4. Based on the conceptual analysis provided above, one can see that Eq.3 correctly attributes causality to c in structures $a1$, $a2$, and $c1$, that is, when the alternative cause a does not possess sufficient power to generate the effect, when the target cause c preempts its competitor a , and when both c and a act in a way called “symmetric overdetermination” in the literature (cf. Hitchcock, 2001; Paul & Hall, 2013; Pearl, 2000; Strevens, 2008). However, the model erroneously attributes causality to c whenever a situation exhibits structure $b2$, in which a preempts c in its efficacy. We believe that preemption of c by preexisting background factors a is the standard case suggested by most cover stories used in causation research (see below for examples).

Cheng and Novick (2005) have considered one additional situation that we have not yet discussed. They developed an equation that allows it to estimate how likely it is

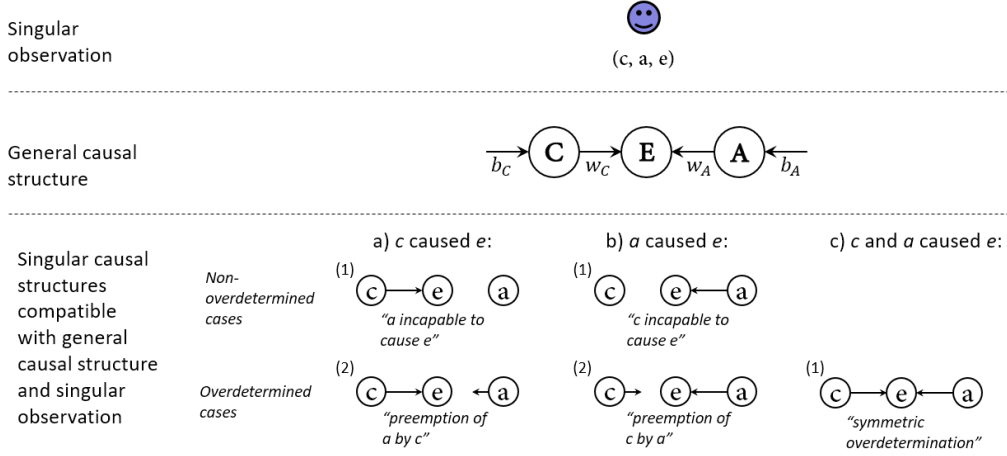


Figure 4. Illustration of the potential singular causal structures in a situation in which c , a , and e were observed that are compatible with the general causal common-effect structure in which C and A are independent probabilistic cause factors of the effect factor E .

that the target cause alone caused the observed occurrence of the effect:

$$P(c \xrightarrow{\text{alone}} e | c, a, e) = \frac{w_C - w_C \cdot w_A}{w_C + w_A - w_C \cdot w_A} = \frac{w_C - w_C \cdot w_A}{P(e | c, a)}. \quad (4)$$

In Eq. 4 the product of w_C and w_A is subtracted from w_C in the numerator, which means that the probabilistically overdetermined cases are now “attributed away” from the target cause. This equation captures the cases described in the contexts $a1$ and $b2$ in Fig 4. Yet, it fails in contexts, such as structure $a2$, in which the target cause preempts the alternative cause in its efficacy.

In sum, Eq. 3 attributes the whole sufficiency overlap to the target cause, and therefore captures $P(c \rightarrow e | c, a, e)$ correctly in situations in which the target cause preempts the alternative cause. Eq. 4, by contrast, attributes the full sufficiency overlap to the alternative cause, and hence captures correctly the probability of the target cause being the singular cause of the effect in situations in which the alternative cause preempts the target cause. The problem of the model is that there is an infinite number of intermediate steps between “the target cause was certainly preempted” and “the

target cause was certainly not preempted.” It cannot handle, for instance, a situation in which a reasoner is uncertain as to whether the target cause was preempted by the alternative cause. Furthermore, the model does not provide an account of the factors influencing intuitions about preemption.

3 A generalization of causal attribution theory sensitive to preemption

To incorporate the possibility of causal preemption, Stephan and Waldmann (2018) have proposed a generalization of the power PC framework of causal attribution. Eq. 5 summarizes the generalized model:

$$P(c \rightarrow e|c, a, e) = \frac{w_C - w_C \cdot w_A \cdot \alpha}{w_C + w_A - w_C \cdot w_A} = \frac{w_C \cdot (1 - w_A \cdot \alpha)}{P(e|c, a)}. \quad (5)$$

Eq. 5 introduces a new parameter, α , which is weighted with the product of w_C and w_A . The product $w_C \cdot w_A \cdot \alpha$ is supposed to capture the probability of preemption of the target cause factor by the competing alternative cause: The intersection of w_C and w_A identifies the sufficiency overlap, i.e., the cases in which C and A are both probabilistically sufficient for E . This part of the term is relevant because the problem of preemption only occurs when the potential causes are both probabilistically sufficient. On these occasions it can either be the case that c preempts a or that a preempts c . Another possibility is that c and a act symmetrically on an occasion, instantiating a situation of “symmetric overdetermination”. The newly introduced α parameter represents an allocation parameter, $0 \leq \alpha \leq 1$, that determines the proportion of the sufficiency overlap in which C is preempted by A . As $w_C \cdot w_A \cdot \alpha$ captures the probability of preemption of the target cause by the alternative causes, it needs to be subtracted from w_C in the numerator.

Consider again the example of the fictitious medical study. We know that $w_C \cdot w_A = \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}$, that is, that there were eight cases in which both potential causes were probabilistically sufficient for the effect. Now imagine a reasoner believed that the alternative cause tends to preempt the target cause with a relatively high, say 75 percent, probability on occasions on which both are probabilistically sufficient. To

model this situation, the α parameter in Eq. 5 needs to be set to 0.75, which yields $w_C \cdot w_A = \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{3}{4} = \frac{1}{4}$ for the product in the numerator. Thus, six of the eight cases of the sufficiency overlap in the example would be attributed away from the target cause, yielding $P(c \rightarrow e|c, a, e) = \frac{1}{2}$. One should therefore conclude that for a randomly sampled singular case in which the effect and the target cause co-occurred, c has caused e with a probability of fifty percent. The two extreme types of situations in which a reasoner is either certain that the target cause preempted the alternative cause or certain that the target cause was preempted by the alternative cause can be modeled by setting the α parameter to 0 or 1, respectively. In the first case, Eq. 5 is identical to Eq.3. In the second case, Eq. 5 reduces to Eq. 4.

Cases of “symmetric overdetermination” can be modeled by setting α to 0. If both potential causes are by definition assumed to act symmetrical in this type of situation, it seems normative to judge both to be singularly causally linked to e (cf. Lagnado & Gerstenberg, 2017; Pearl, 2000; Schaffer, 2003). Empirical studies that explicitly investigated this type of situation have yielded mixed results (see Sloman & Lagnado, 2015, for an overview). Some studies (Spellman & Kincannon, 2001) found that people seem to regard symmetrically overdetermining causes as highly causal. Others found that such cases are viewed as less causal than cases of preemption (e.g., Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015).

The experiments conducted by Stephan and Waldmann (2018) aimed to provide a first demonstration that subjects’ singular causation judgments deviate substantially from the predictions of the standard causal attribution model when the context suggests that the alternative causes tend to preempt the target cause in its efficacy. In their Experiment 1, subjects were presented with a causal induction task in which the contingency information about a single cause and effect was presented in a summary format (cf. Fig.,1). Subjects were then asked to make a singular causation judgment about a randomly selected case from the treatment group that exhibited the effect.

Stephan and Waldmann (2018) used a cover story in which it was plausible that the alternative causes preempted the target cause in its efficacy. The cover story

described a fictitious scientific study (cf. Griffiths & Tenenbaum, 2005) in which biologists investigated the influence of a chemical substance on the expression of a particular gene in mice. It was described that the biologists examined two random samples of mice, one treated with the chemical substance and one that served as a control group. It was explicitly stated that the control group had to be included because it is known that some mice might be expressing the gene for natural reasons. This information was assumed to induce the intuition that preemption of the target cause by alternative causes was likely. It seems plausible that mice who were expressing the gene for natural reasons were already doing so before the biologists selected and treated the mice with the chemical substance. After subjects had inspected the contingency information, they were shown a single mouse from the treatment group that exhibited the effect and were asked to indicate how confident they were that this particular mouse was exhibiting the effect because of the treatment. Stephan and Waldmann (2018) therefore modeled the studied cases assuming an α value of 1. The experiments supported this model. Additional control conditions indicated that subjects did indeed assume high values for the α parameter of the extended model.

Although subjects' singular causation judgments in the experiment were captured well by the extended model incorporating the possibility of preemption, they are also in line with the predictions that the standard model makes for situations in which the goal is to estimate the probability with which the target cause alone produced the effect (cf. Eq. 5). A shortcoming of the study was that the factors that influence assumptions about the α parameter were neither spelled out nor experimentally manipulated.

4 Assessments of singular causation are sensitive to temporal assumptions

We have argued in the previous section that the assessment of singular causation needs to take the possibility of causal preemption into account. We showed how the causal power theory of causal attribution can be extended in order to handle the problem of causal preemption by incorporating the product of w_C , w_A , and the newly introduced parameter α . A crucial component for the estimation of the probability of

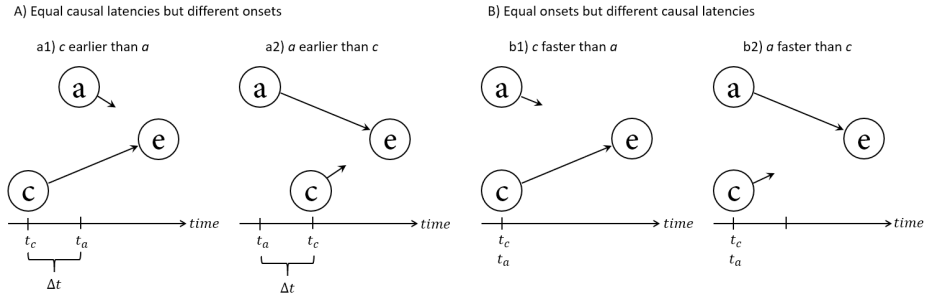


Figure 5. Illustration of the temporal components relevant for the assessment of the potential preemptive relation between two potential causes of an effect. A) illustrates the relevance of the instantiation times of the potential causes. B) illustrates the relevance of the causal latencies of the potential causes.

preemption is information about the causal strengths of the potential causes: It needs to be determined in a first step how likely it is that the effect was probabilistically overdetermined by the potential causes on the target occasion because causal preemption is only in these occasions possible. In a second step, it needs to be estimated how likely it is that the target cause was actually preempted by the alternative causes on such occasions. This probability is expressed by the newly introduced α parameter.

Intuitions about causal preemption involve considerations about the temporal relations between the potential causes and the effect. We propose that temporal assumptions about these relations determine the value of α . To demonstrate the plausibility of this claim, we will again focus on situations in which there are two potential causes of an effect E , the target cause C and an alternative cause A .

4.1 Onset differences

One relevant type of temporal information is information about the onset times of the potential causes. Everything else being equal, if a occurs earlier than c on an occasion on which both are probabilistically sufficient for the effect, then a will preempt c . Conversely, c preempts a when c occurs earlier than a in this kind of situation. The standard example about Billy and Suzy discussed above exemplifies this structure. The instantiation times of two potential causes c and a can be denoted as t_c and t_a , and the

difference between them can be denoted as $\Delta_t = t_a - t_c$. A schematic illustration using time indexed neuron diagrams is depicted in Fig. 5A. Formally, if two potential causes c and a are both sufficiently powerful on an occasion, c preempts a in causing e when Δ_t is positive. Conversely, a preempts c in causing e when Δ_t is negative. A general-level representation Δ_T can be obtained through averaging over the observations of Δ_t .

4.2 Causal latencies

Another type of temporal information relevant for the assessment of preemption is information about the causal latencies of the potential causes. By causal latency we mean the time it takes a cause to produce its effect. For example, swallowing a dose of a pain killer does not bring relief right away. It takes a particular amount of time until the medicine manifests its capacity. When we enter an elevator and press a button, it takes a moment until the elevator begins to move. Even when we press a button on the remote control, the TV does not react instantaneously.

An important characteristic of causal latencies is that they are, like causal powers, unobservable properties of causes. The causal latency of a cause factor C , which can be formally expressed as $t_{(C \rightarrow E)}$ (cf. Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018), must be inferred from observable onset differences between singular instantiations of C and E . Only in an environment that is shielded from potential alternative causes does the observed onset difference between c and e represent a direct measure of the causal latency of c ; just like Δ_P is only representing a direct measure of causal power when the base rate of the effect in the learning context is zero. The remote control and the pain killer scenarios are good illustrations for this property. Whereas TVs rarely turn on unless we press the button on the remote control, headaches typically subside even without the help of aspirin. Both the causal power as well as the causal latency of the painkiller are much harder to learn than the causal power and the causal latency of the remote control.

Neuron diagrams illustrating the relevance of causal latency for preemption are depicted in Fig. 5B. Everything else being equal, a probabilistically sufficient potential

cause c of the effect e preempts a simultaneously sufficient potential alternative cause a on an occasion if c 's causal latency, $t_{(c \rightarrow e)}$, is smaller than a 's, $t_{(a \rightarrow e)}$. Causal latency here refers to the time it would have taken the target cause to produce the effect if it had acted in isolation. Conversely, a preempts c in causing e if a 's causal latency is smaller than c 's on such an occasion. Of course, the “everything else being equal” criterion will rarely hold in real life situations. On most occasions different onset times will be combined with different causal latencies. This leads to the possibility that causes with high causal latency can even preempt alternative causes if they have an onset advantage on their side. Likewise, causes that occur after their competitors can still manage to preempt them if they have a small causal latency.

The example about the elevator or about the light traveling from the remote control to the receiver shows that it might sometimes be appropriate to represent the causal latency of a cause factor as a stable, fixed value. In most contexts, however, causal latencies will appear variable. The pain killer scenario provides an example for such a context. A dose of medication might be expected to take effect after about twenty minutes, but one would probably not be too surprised if the effect occurred a bit earlier or a bit later. In such contexts it seems appropriate to represent the causal latency of a factor as a random variable following a particular distribution. Variations of causal latencies might be causally explained by the operation of unobserved hasteners or delayers (Lagnado & Speekenbrink, 2010). For example, the time it takes an Aspirin to bring relief is influenced by various factors, including the momentary concentration of gastric acid in the person's stomach, her momentary blood pressure, and so on.

A standard way of modeling latencies that exhibit variation is to use gamma distributions, which are, for example, used in queueing theory to model waiting times (Shortle, Thompson, Gross, & Harris, 2018). In a recent study in which Bramley et al. (2018) investigated the role of time in general-causal structure induction, gamma distributions were used to model causal latencies. The gamma distribution belongs to the family of continuous probability distributions and represents a generalization of the exponential distribution. It is characterized by two parameters: shape, $\kappa > 0$, and scale,

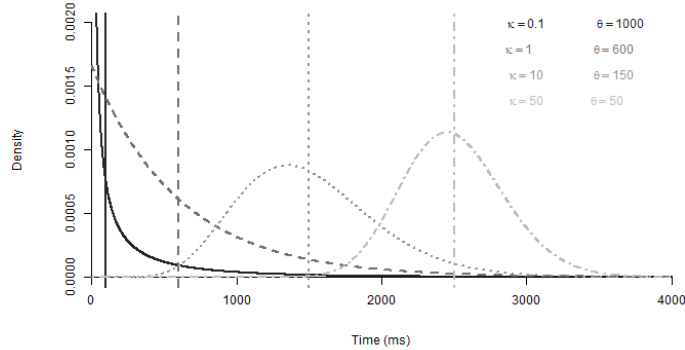


Figure 6. Example of gamma distributions with different shape (κ) and scale (θ) parameters. The expected value of a gamma distribution is given by $E[X] = \kappa \cdot \theta$. The expected values of the example distributions are illustrated by the vertical lines.

$\theta > 0$. The expected value of a random variable X following a gamma distribution is $E[X] = \kappa \cdot \theta$. Its variance is $Var[X] = \kappa \cdot \theta^2$. Different gamma distributions together with their expected values are shown in Fig. 6.

4.3 Modeling the factors underlying the α parameter

The formalization of the different temporal components that influence the probability of preemption allow it to quantify α for different types of situations. We will focus again on the type of situation that we have already considered above in which a reasoner has learned that two potential causes C and A and their effect E have co-occurred (c, a, e) on an occasion. We here consider contexts in which the causal latencies of the potential causes have been learned through multiple observations and in which the causes' onset difference in the target situation is known. In these cases, the probability that the target cause c was preempted by the alternative cause a if both happen to be simultaneously sufficient for the effect corresponds to:

$$\alpha = P(t_{a \rightarrow e} + \Delta_t < t_{c \rightarrow e} | e, c, a). \quad (6)$$

In this situation, α is the probability that the sum of the causal latency of the

competing cause a and the time lag between c 's and a 's onset times is smaller than the causal latency of c given e , c , a . When the causal latencies can be represented as fixed values, like in the remote control example, α can be determined by comparing the simple sum of $t_{(a \rightarrow e)}$ and Δ_t with $t_{(c \rightarrow e)}$. In this case, α will either be 0 or 1. In the more complicated case in which the causal latencies exhibit variability and are modeled with gamma distributions, α can take on any value between 0 and 1, and can be estimated with the following Monte Carlo (MC) algorithm:

1. Sample N pairs of causal latencies $(t_{c \rightarrow e}, t_{a \rightarrow e})$ from the latency distributions $(t_{C \rightarrow E}, t_{A \rightarrow E})$ of C and A , respectively.
2. Calculate $t_{a' \rightarrow e} = t_{a \rightarrow e} + \Delta_t$ for all sampled $t_{a \rightarrow e}$ -values.
3. Count all pairs for which $t_{a' \rightarrow e} < t_{c \rightarrow e}$.
4. Divide this count by N .

For an illustration, consider Fig. 6 again. Assume that the causal latency of the target cause C followed the gamma distribution with the parameters $\kappa = 10$ and $\theta = 150$, and that the causal latency of the alternative cause A corresponded to the distribution with the parameters $\kappa = 50$ and $\theta = 50$. The expected value of the target cause's causal latency is $E = \kappa \cdot \theta = 1500 \text{ ms}$, while the causal latency of the alternative cause has an expected value of $E = \kappa \cdot \theta = 2500 \text{ ms}$. Hence, the target cause tends to manifest its causal capacity quicker than its competitor. However, there is also a substantial overlap between the two causal latencies which implies that there will be occasions in which the alternative cause outperforms the target cause. As we have seen, the α parameter captures the probability with which this happens by additionally taking the onset asynchrony of the causes into account. For occasions on which the two causes occur simultaneously (i.e., $\Delta_t = 0$) the MC algorithm presented above yields a value of $\alpha = 0.05$, which means that the probability that a acts quicker than c in such a singular occasion is five percent. Now imagine that it is known that both causes not only occurred at the same time but also that their causal latencies

followed the same gamma distribution, that is, that the two distributions fully overlap. In this case, α would be 0.50, because in a set of randomly sampled pairs of causal latencies, the latency of the alternative cause would be smaller in half of the pairs.

Finally, consider the computationally simpler case in which causal latencies can be represented as fixed values, like in the remote control example. The same sampling algorithm can be used to estimate α in this case. However, it is easy to see that α will either be 1 or 0 in such situations depending on whether the sum of $t_{(a \rightarrow e)}$ and Δ_t is smaller than $t_{(c \rightarrow e)}$ or not.

The focus of the present research is on the role of temporal assumptions in singular causation judgments. Given, however, that singular and general causation are coupled it is important to know what role temporal assumptions play in general causal learning (see Buehner, 2017, for a review). A typical early finding was that reasoners seem to have a strong preference for spatio-temporal contiguity when learning about the strengths of causes (e.g., Shanks, Pearson, & Dickinson, 1989). These studies manipulated temporal gaps and found that learning about contingencies becomes more difficult with increasing spatio-temporal distance between causes and effects.

However, contiguity becomes less potent once background knowledge about latencies is activated by means of cover stories drawing on prior knowledge. For example, Haggmayer and Waldmann (2002) showed that assumptions about temporal delays determine which events in a sequence of observed events are considered candidates for a causal relation, which in turn affected which contingencies were learned as indicators of causal strength.

Another study by Buehner and McGregor (2006) confronted subjects with an unfamiliar apparatus in which marbles running down a pathway could activate different switches connected to a light bulb. During an inspection phase, the steepness of the pathways within the machine was varied, which either suggested a relatively small or a relatively large delay between marble insertion and the activation of light. During the test phase, the mechanism of the device was covered and subjects observed multiple instances of marble insertion and light activation. The observed delays were either

compatible or incompatible with previously learned causal mechanisms. The results showed that subjects who observed highly contiguous successions of events only gave high causality ratings when the mechanism in fact suggested short delays, while causality ratings were low when subjects expected the delays to be longer.

In another study, Lagnado and Speekenbrink (2010) showed that lowered causal strength judgments with increasing causal delays might be due to subjects' assumptions about the preemptive relation between the target cause and potential alternative causes that occur in the delay period. In their experiment, Lagnado and Speekenbrink (2010) varied the causal latency of the target cause factor and the probability with which alternative causes of the effect occurred during the cause-effect interval. This experimental design allowed it to compare causal judgments between conditions with equal causal latencies but varying probabilities with which potential alternative causes occurred before the effect occurred. Causal judgments tended to be lower when the probability of the occurrence of alternative causes before the occurrence of the effect was high but not when this probability was low, presumably because the occurrence of alternative causes before the occurrence of the effect opens up the possibility that the target cause was preempted. Variation in causal latency alone did not affect causal judgments in this study.

Whereas the study of Lagnado and Speekenbrink (2010) focused on the role of temporal relations and possible cases of preemption in samples of causal events (i.e., general causation learning task), our focus in the following four experiments will be on the interaction between causal strength, temporal assumptions and preemption in judgments about singular causation relations. These studies will test our generalized model of singular causation judgments.

5 Experiment 1

The goal of Experiment 1 was to test whether the two temporal components, causal latency and onset difference, would actually affect reasoners' singular causation judgments in the way predicted by the new model. The α parameter of the new model

Table 1

Overview of the set of different situations tested in Experiment 1.

	(1) Different onsets but equal causal latencies		(2) Equal onsets but different causal latencies		(3) Different onsets and causal latencies (A)		(4) Different onsets and causal latencies (B)	
	Cause A	Cause B	Cause A	Cause B	Cause A	Cause B	Cause A	Cause B
Onset (ms)	800	1600	800	800	1600	800	800	2400
Causal latency (ms)	800	800	800	1600	800	2400	1600	800
Sum	1600	2400	1600	2400	2400	3200	2400	3200

integrates onset and causal latency information and assigns equally important roles to both factors, which means that a relatively slowly acting cause might compensate its disadvantage on this dimension with an advantage on the onset dimension and vice versa. This implies, for example, that reasoners should sometimes attribute causality to the cause that had a larger temporal distance to the effect on a singular occasion than the alternative cause. The present experiment tested if subjects would actually integrate both types of information in this way. Moreover, the first study aimed to isolate the influence of the temporal factors from the influence of causal power. This was achieved by testing causes with deterministic powers. The standard causal power model of attribution (Eq. 3) predicts high singular causation ratings in this case. Finally, to keep things relatively simple in this first experiment, the cause factors were assigned fixed, non-varying causal latencies.

The four different scenarios involving two potential causes, A and B, of a single effect are shown in Table 1. It can be seen that the two potential causes A and B were associated with different combinations of causal latency and onset times across the different situations. When both potential causes have equal causal powers, our model predicts that subjects should give high singular causation ratings for to the cause that minimizes the sum of causal latency and onset time, as it is unlikely that this cause was preempted by its competitor. The sum of causal latency and onset time for each of the two potential causes is listed in the last row of Tab. 1. Subjects were expected to give high singular causation judgments when the target cause was Cause A and low singular causation judgments when the target cause was Cause B.

In the first scenario the influence of onset differences was isolated and both causes had identical causal latencies. While both potential causes always produced the effect after 800 ms in this scenario, the onset difference was $\Delta_t = 800ms$. The second scenario tested the influence of causal latencies. Both causes always occurred simultaneously 800ms after a trial had begun, but produced the effect with different latencies.

The last two scenarios (3 and 4) pitted causal latency and onset difference against each other. Two different versions (A and B) tested whether subjects actually integrated both temporal components or whether they considered one component more important than the other. For example, had we only tested the third scenario, higher singular causation judgments for cause A could have been explained by a bias towards relatively low causal latencies. Likewise, results in the expected direction in scenario 4 could have been explained by a bias towards causes with an onset advantage.

5.1 Method

5.1.1 Participants. Three hundred and eighty-four subjects ($M_{age} = 35.14$, $SD_{age} = 12.38$, 202 female, 178 male, four subjects did neither indicate male nor female as their gender) were recruited via Prolific (www.prolific.ac). This sample size ($n = 96$ subjects per scenario) allows it to detect medium effects (of $d = 0.60$) in single group comparisons conducted with conservative two-sided t-tests with more than 80 percent test-power. We planned to analyze the data using planned contrasts, however, which guarantees an even higher test-power. No data were analyzed until $n = 96$ subjects per condition had completed the study. Subjects were at least 18 years old, had at least an A-level degree, were native English speakers, and had a work acceptance rate of at least 90 percent. Subjects were paid £ 1.25 for their participation.

5.1.2 Design, materials, and procedure. Subjects were randomly assigned to one of four different scenarios ($n = 96$) that varied with respect to the combination of causal latencies and onset differences of the two potential causes as shown in Tab. 1. Several factors were balanced between subjects that will be described below.

Subjects were presented with a story about a fictitious medieval kingdom called

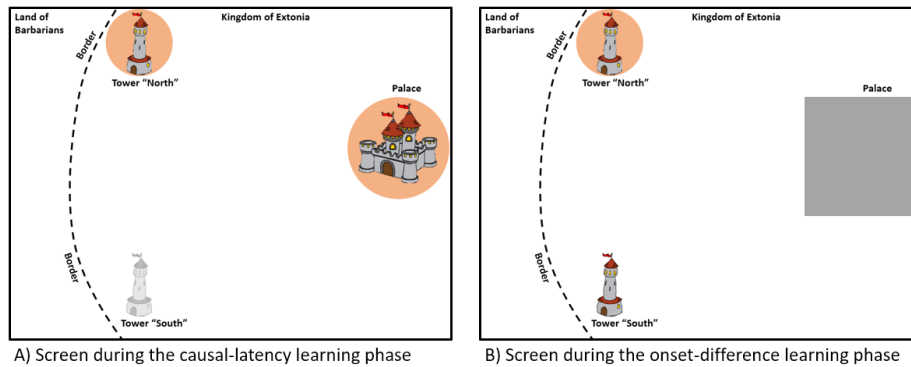


Figure 7. Illustration of the learning task used in Experiment 1.

“Extonia” whose king had two watchtowers (“North” and “South”) built at the border to protect the empire from invading barbarians. Subjects learned that the watchtowers were instructed to send carrier pigeons to the palace to cause alarm whenever they spotted barbarians approaching the border. It was described that pigeons automatically set off an alarm the moment they arrive at the palace. Participants were then asked to take the perspective of Extonia’s secretary of defense who wants to inspect the efficacy of the kingdom’s defense system. It was mentioned that two factors must be investigated: the flight durations of the pigeons from the two towers and the alertness of each tower (i.e., how quickly each tower reacts when barbarians are coming). The flight durations of the pigeons constituted our manipulation of the towers’ causal latencies, whereas the reaction times of the towers allowed us to manipulate the onset differences.

Participants were further instructed that they will learn about the two factors in separate learning phases. They were also informed that they will observe ten pigeons from each tower in the flight-duration phase (i.e., 20 observations in total) and that they will make ten observations in the alertness phase. The instructions additionally contained illustrations of how the animations in the two different learning phases will look (see Fig. 7). Finally, participants were informed that at the end they will be asked to answer a causal query about a singular occasion on which the palace was alarmed. Before participants could proceed to the learning phase, they had to pass five instruction check questions probing their understanding of the scenario. Subjects who

failed to answer all these questions correctly four times in a row were counted as invalid and excluded from the data set prior to any analysis of the data.

Whether subjects first learned the flight durations of the pigeons (cf. Fig. 7A) or began with the onset differences (cf. Fig. 7B) between the two towers was balanced between subjects. The flight-duration learning phase consisted of two parts, as subjects observed the two towers separately. Whether participants began with tower “North” or “South” was also counterbalanced between subjects. In each trial, the sending of a carrier pigeon was indicated by a colored circle surrounding the active watchtower with a delay of $500ms$. The corresponding inactive tower was transparent (see Fig. 7A). The arrival of the pigeon at the palace was indicated by a colored circle surrounding the palace. The circles always remained visible until a trial had ended. The ten flight durations presented for each tower depended on the causal latencies tested in the respective conditions. Whether tower “North” or tower “South” had the faster pigeons was counterbalanced between subjects. After each trial participants had to click a “Next” button to proceed to the next observation. The “Next” button was operational $500ms$ after the circle around the palace was shown. After subjects finished their observations, they proceeded to an intermediate screen on which they were informed that they will now investigate the second relevant factor. The content of the intermediate screen depended on the order of the learning phases. The animation shown during the onset-learning phase resembled the illustration depicted in Fig. 7B. The palace was masked with a grey rectangle in this phase to underline that it was irrelevant here. As in the flight-duration phase, the sending of letter pigeons was indicated by colored circles surrounding the towers. The ten onset-differences that subjects observed in this phase depended on condition. Whether tower “North” or tower “South” reacted quicker was balanced between subjects.

Subjects then proceeded to the description of the target situation. A singular occasion was described on which the palace had been alarmed and on which a horde of invading barbarians had been repelled successfully. Participants read that the people of Extonia wanted to decorate the crew of the tower that caused the alarm on this

occasion. All that the Extonians knew, however, was that both towers had spotted the barbarians and had sent a pigeon to the palace on this occasion. Subjects were also presented with an image showing colored circles around each tower and around the palace. We then asked subjects to indicate how strongly they believed that the alarm was caused by tower “North”/ “South” on this occasion. Judgments were provided using an eleven-point rating scale with end points “definitely not caused by Tower ‘North/ South’” and “definitely caused by Tower ‘North/ South’” (the midpoint was labeled “50:50”). This rating was intended to measure $P(c \rightarrow e|c, a, e)$. Whether the test question referred to tower “North” or “South” was counterbalanced between subjects. A bipolar rating scale with one tower on either side of the scale would have forced subjects to decide for one of the towers, which would have introduced a disadvantage for the power PC model of causal attribution. As the power PC model of causal attribution solely focuses on causal power information, it predicts that both towers should be considered causal in this case. Finally, the orientation of the rating scale (i.e., whether high confidence had to be expressed on the left or on the right side of the scale) was also counterbalanced between subjects.

5.2 Results and discussion

The results are summarized in Fig. 8. As can be seen, subjects made different singular causation judgments for the two potential causes in the four scenarios, which shows that they applied their temporal background knowledge about both the onset differences and the causal latencies. The upper two panels in Fig. 8 show the results for the two scenarios that tested each temporal component in isolation (Scenarios 1 and 2 in Tab. 1). The left panel shows the results for Scenario 1 in which both causes had the same causal latency but different onsets, while the right panel shows the results for Scenario 2 in which both causes had identical onsets but different causal latencies. In the condition in which one cause factor tended to occur earlier than the other but both had identical causal latencies (Scenario 1) subjects who were asked about the early cause (Cause A in Table 1) gave higher singular causation judgments ($M = 0.75$,

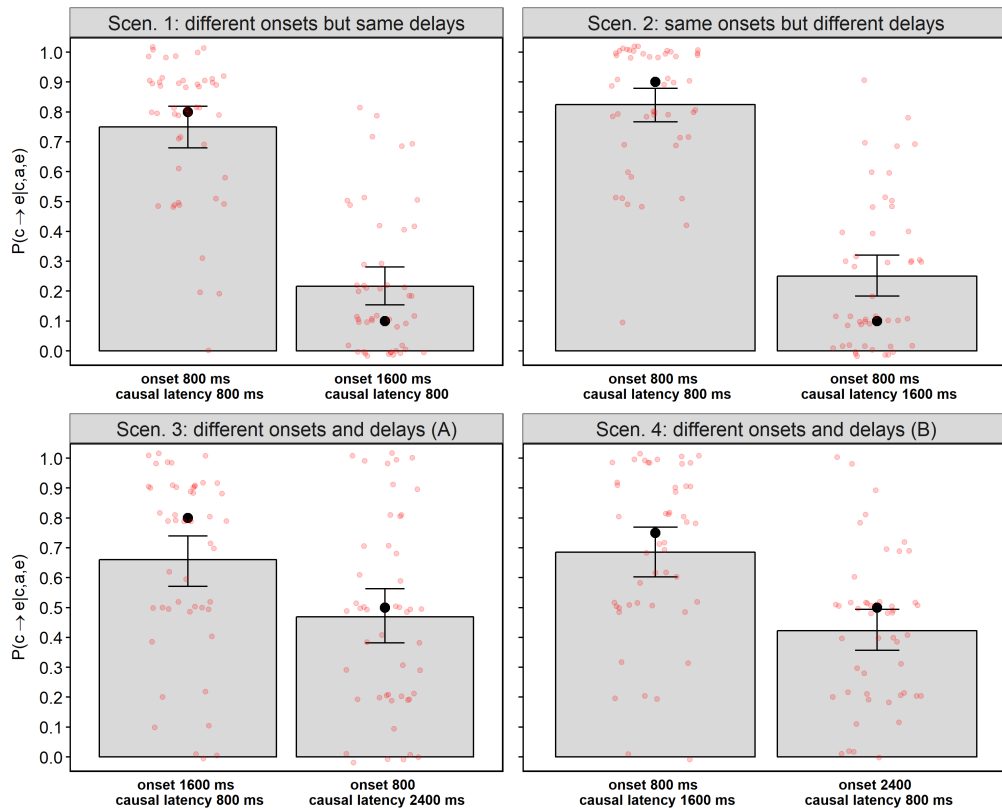


Figure 8. Results (bars represent mean singular causation ratings; error bars denote 95% bootstrapped *CI*s; red points show jittered individual ratings; black points show medians) of Experiment 1.

$SD = 0.24$, $Mdn = 0.80$, $95\% CI = 0.07$) than subjects who were asked about the “late” cause ($M = 0.21$, $SD = 0.24$, $Mdn = 0.10$, $95\% CI = 0.07$). A planned contrast revealed that this difference was significant, $t(376) = 9.9$, $p < .0001$, $d = 2.25$. In the scenario in which both causes had synchronous onset times but varying causal latencies (Scenario 2), subjects gave higher singular causation ratings when asked about the cause with the smaller causal latency ($M = 0.83$, $SD = 0.21$, $Mdn = 0.90$, $95\% CI = 0.06$) than when asked about the cause with the higher causal latency ($M = 0.25$, $SD = 0.25$, $Mdn = 0.10$, $95\% CI = 0.07$). A planned contrast revealed that this difference was significant, $t(376) = 10.68$, $p < .0001$, $d = 2.76$.

The results for the first two scenarios show that subjects used their knowledge

about causal latencies and onset differences when making singular causation judgments. To test whether subjects actually integrated the two types of temporal information as predicted by the model, ratings in Scenarios 3 and 4 were compared. These ratings are shown in the bottom panels of Figure 8. They reveal that subjects tended to answer as predicted by the model. In Scenario 3, subjects gave higher singular causation ratings when asked about the cause with the smaller causal latency ($M = 0.66$, $SD = 0.30$, $Mdn = 0.80$, $95\% CI = 0.09$) than when asked about the cause with the larger causal latency ($M = 0.47$, $SD = 0.32$, $Mdn = 0.50$, $95\% CI = 0.09$), even though the onset difference pointed in the other direction. A planned contrast testing this difference was significant, $t(376) = 3.56$, $p < .001$, $d = 0.63$. However, one possible explanation for this pattern is that subjects might have had a general bias towards smaller causal latencies. The ratings obtained in Scenario 4 show that this does not explain the judgments. Subjects here gave higher singular causation ratings when asked about the cause with the larger causal latency ($M = 0.69$, $SD = 0.28$, $Mdn = 0.75$, $95\% CI = 0.08$) than when asked about the cause with the smaller causal latency ($M = 0.42$, $SD = 0.25$, $Mdn = 0.50$, $95\% CI = 0.07$). A planned contrast testing this difference was significant, $t(376) = 4.87$, $p < .0001$, $d = 0.96$. Subjects did also not have a general tendency to favor the cause that had an onset advantage, as one could have hypothesized based on the results of Scenario 4 alone. In Scenario 3, the cause with the onset advantage received lower ratings than the cause with the onset disadvantage. Additional evidence for subjects considering both temporal dimensions to be equally important is that the difference between the two causes is similar across the Scenarios 3 and 4.

The results of this experiment show that subjects integrated both types of information when making singular causation judgments, and that they, in line with what the new model predicts, generally favored the cause that minimized the sum of Δ_t and causal latency. The fact that the observed differences were overall smaller in Scenarios 3 and 4 than in Scenarios 1 and 2, where the causes differed on only one temporal dimension, can be explained by the higher difficulty of the computation in the more complex cases. The more complex conditions were also more memory demanding.

Subjects had to retrieve from memory information about the degree to which the causes differed on both dimensions, while it sufficed in the first two scenarios to remember the ordinal ranking of the causes on a single discriminating dimension.

6 Experiment 2

Experiment 1 represents a successful demonstration that reasoners apply more than just their knowledge about the causal powers of the potential causes when assessing how likely it is that a particular factor C was causal on a singular occasion. We demonstrated that reasoners also rely on what they have learned about different types of temporal factors, the onset times of the causes and their causal latencies.

The causal latencies that subjects learned in Experiment 1 had fixed non-varying values, which is something that people might only rarely experience in their everyday lives. As we have pointed out above, the more natural case is that causal latencies exhibit some degree of variability. The goal of the second experiment was therefore to investigate such cases. We compared different conditions in which the two potential causes were paired with causal latencies that followed different gamma distributions. On the assumption that variations of causal latencies are the default situation, we expected subjects to be sensitive to this factor. We expected that subjects' singular causation judgments will be in line with the predictions of our modified model. However, another possibility is that subjects might only estimate the expected values of the causal latencies. The way in which we manipulated causal latencies in the present study allowed us to test both possibilities (see Fig. 10). Again, we restricted the scenarios to deterministic causes to be able to focus on the contribution of causal latency information. Furthermore, subjects were instructed that the onsets of both causes occurred simultaneously in each scenario.

The pairs of gamma distributions that were contrasted in the five different conditions of the experiment are shown in Fig. 9. As can be seen, the distances between the distributions (G1 - G5; higher numbers indicate higher expected values) in the different conditions varied quantitatively, from fairly extreme in Condition 1, for

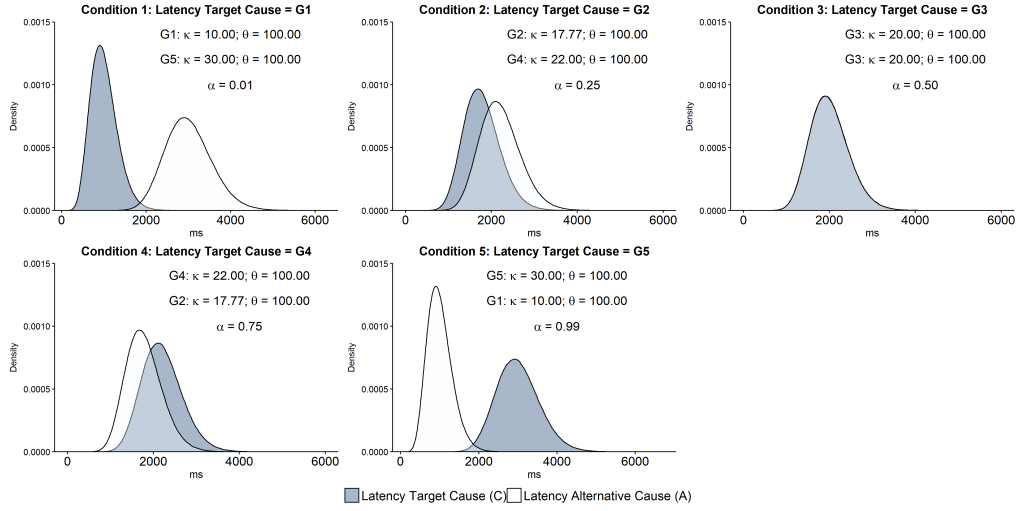


Figure 9. Pairs of gamma distributions contrasted in the five conditions of Exp. 1. The shape (κ) and scale (θ) parameters of the five different gamma distributions are listed for each pair. The depicted α values were obtained using the MC algorithm corresponding to Eq. 6. The dark distributions show the causal latencies of the target cause and the light distributions show the causal latencies of the alternative cause.

example, to identical in Condition 3. The latency distribution belonging to the target cause C in each condition is depicted in dark blue. In Condition 1, for example, which contrasted the distributions G1 and G5, the target cause’s latency follows G1, whereas the latency of the alternative cause follows G5. Conditions 4 and 5 involve the same pairs of distributions as Conditions 1 and 2, but the distribution associated with the target cause was changed. Fig. 9 also shows the different α values estimated with the sampling algorithm presented above. For the first pair, for example, α equals 0.01. In the case of deterministic causes α corresponds to the probability that c was preempted by a . Thus, participants should be confident in this condition that it was indeed the target cause c that brought about the observed outcome. In the fifth condition, by contrast, in which C follows G5 and A follows G1, participants should be confident that c did not cause the effect. Fig. 9 also shows that all the other conditions should elicit more uncertainty. In the third condition, for example, in which C and A have the same latency distribution (G3), $\alpha = 0.50$, participants should be maximally uncertain about

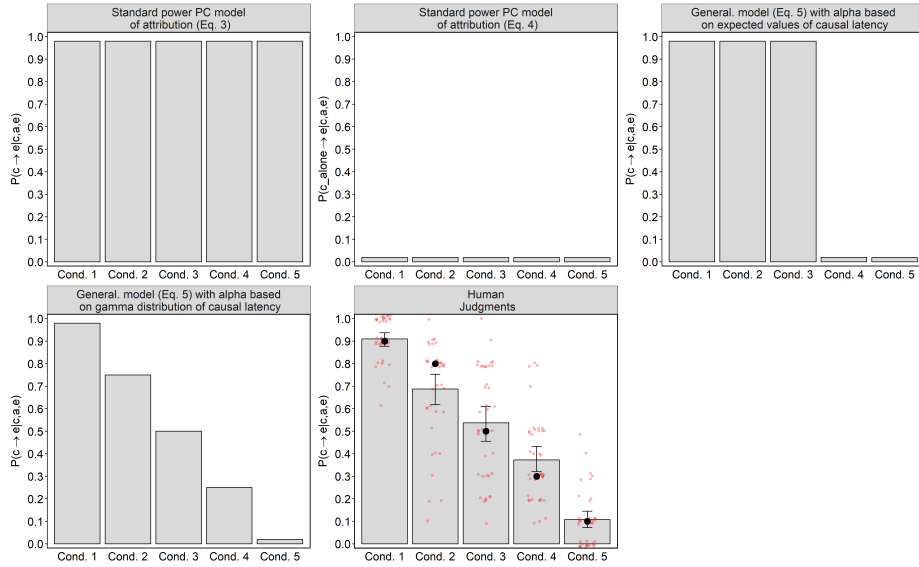


Figure 10. Model predictions and results (bars show mean singular causation ratings; error bars denote 95% bootstrapped *CI*s; red points show jittered individual ratings; black points show medians) for the different conditions (see Fig. 9) of Exp. 2. The predictions of the Standard power PC model of causal attribution were obtained from Equations. 3 and 4. The predictions of the generalized model were based on different ways of estimating alpha. The first of these plots shows the predictions obtained when α is calculated based on the expected values of the respective gamma distributions. The second plot shows the predictions obtained when α is estimated based on the gamma distributions.

the singular cause of the outcome.

Fig. 10 shows the predictions that the different models make. The predictions made by the standard power PC model of causal attribution are shown in the first two panels. The first panel shows the predictions derived from Eq. 3. We also included the predictions derived from Eq. 4f, which are shown in the second panel. We wanted to show that subjects do not misinterpret our test question and understand it as one that is asking for the probability that “c alone” caused e on the target occasion. Note that the “c alone” query according to the standard model asks for the probability with which c alone was probabilistically sufficient on an occasion; the standard model equates

sufficiency with causality. If subjects equated sufficiency with causality and understood the test query as a “c alone” query, we should expect low ratings in all conditions (see second panel in Fig. 10). The third panel in Fig. 10 shows the predictions that the generalized model makes when α is estimated based on the expected values of the causes’ causal latencies. In this case, the cause factor with the lower expected value of causal latency is identified as the singular cause. When both causes have the same expected causal latency value (Condition 3), the situation is treated as a case of symmetric overdetermination. Finally, the fourth panel shows the predictions that the new model (Eq. 5) makes when α is estimated based on the respective causal latency distributions.

6.1 Methods

6.1.1 Participants. Two hundred subjects ($M_{age} = 27.14$, $SD_{age} = 8.71$, 119 female, 91 male) were recruited via Prolific (www.prolific.ac). This sample size ($n = 40$ subjects per condition) allows it to detect medium effects (of $d = 0.60$) in simple group comparisons using conservative t-tests with more than 80 percent test power. The plan, however, was to analyze the data by fitting quantitative trends, which guarantees an even larger test power for medium effects. No data were analyzed until $n = 40$ subjects per condition had completed the study. Subjects were at least 18 years old, had at least an A-level degree, were native English speakers, and had a work acceptance rate of at least 90 percent. Subjects were paid £ 0.80 for their participation.

6.1.2 Design, materials, and procedure. Participants were randomly assigned to one of five conditions, which varied with respect to the contrasted gamma distributions and with respect to the gamma distribution associated with the target cause (see Fig. 9). Because we included several balancing factors, the full design had 40 conditions. The additional balancing factors will be introduced below.

The cover story and the paradigm were largely identical with the one in Experiment 1. Subjects read that Extonia’s secretary of defense wanted to inspect the flight durations of the pigeons from the two towers. Then subjects were instructed that

the flight durations tend to differ between different pigeons, and that subjects will therefore observe a sample of thirteen pigeons from each tower. Subjects were also informed that they will be asked to answer a causal query at the end about a singular instance in which the palace was alarmed. Before participants could proceed to the learning task, they had to pass five instruction check questions testing their understanding of the scenario. Subjects who failed to answer all these questions correctly four times in a row were counted as invalid and excluded from the data set prior to any analyses of the data.

During the causal-latency learning task the screen looked similar to the picture shown in Fig. 7A. Whether participants began their observations with tower “North” or “South” was counterbalanced between subjects, like in Experiment 1. The thirteen flight durations presented for each tower corresponded to thirteen quantiles of the respective gamma distribution. This was done so that relatively small but yet representative samples of the respective distributions could be presented. After each trial, participants had to click a “Next” button, which was operational 500ms after the circle around the palace had been displayed. The thirteen flight durations per tower were presented in random order. After subjects had observed both towers, they proceeded to an intermediate screen on which they were informed that they will next be shown the singular event to which the test query refers.

Like in Experiment 1, the test scenario described a singular event in which the palace was alarmed and a horde of invading barbarians could be repelled successfully. Participants read that the people of Extonia wanted to decorate the crew of the tower that caused the alarm and that it was known that both towers had sent their pigeons simultaneously. Subjects had to indicate how strongly they believed that the alarm was caused by tower “North”/ “South”. Like in Experiment 1, judgments were provided using an eleven-point rating scale with the end points “definitely not caused by Tower ‘North/ South” and “definitely caused by Tower ‘North/ South” (the midpoint was labeled “50:50”). Whether the test question referred to tower “North” or tower “South” was counterbalanced between subjects. The orientation of the rating scale (i.e., whether

Table 2

Summary of the results of Exp 2.

	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
Median	0.90	0.80	0.50	0.30	0.10
Mean	0.91	0.69	0.54	0.37	0.11
<i>SD</i>	0.10	0.22	0.24	0.19	0.12
95% <i>CI</i>	0.03	0.07	0.08	0.06	0.04

high confidence had to be expressed on the left or on the right side of the scale) was also counterbalanced between subjects.

6.2 Results and Discussion

The results are summarized in Fig. 10 and in Tab.2. As can be seen, participants' singular causation ratings followed a negative linear trend. A polynomial trend analysis revealed that the observed negative linear trend was significant, $F(4, 195) = 111.70$, $p < .001$, $r = .83$. No other polynomial trend was significant.

The results are at odds with the predictions made by the two variants of the standard power PC model of causal attribution that we considered (Equations 3 and 4). As had already been observed in Experiment 1, subjects used information about the causal latencies of the potential causes to derive singular causation judgments. Moreover, the singular causation judgments closely followed the predictions of the version of our new model that utilizes gamma distributions to represent causal latencies (Eq. 5). There was a high correlation between model predictions based on the gamma distributions and individual ratings, $r = .83$, $p < .001$, which is depicted graphically in the left scatterplot of Fig. 11. We also analyzed the fit between model predictions and mean singular causation judgments in the different conditions. The correlation between model predictions and group means is depicted in the right panel of Fig. 11. This correlation amounted to $r = .99$, $p < .001$.

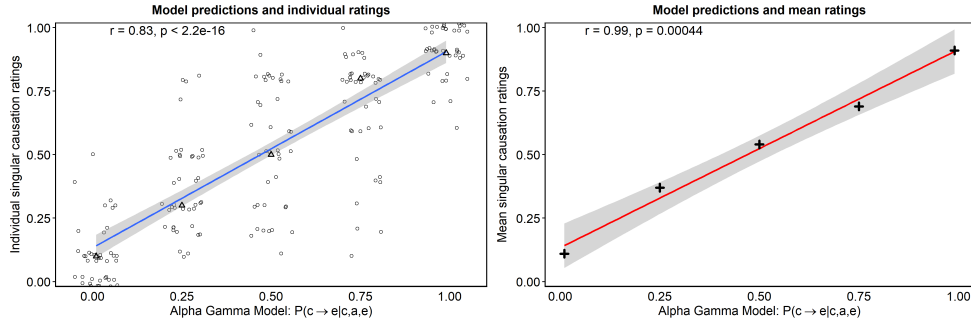


Figure 11. Scatterplots depicting the correlations between predictions of the generalized model and individual singular causation judgments (left; triangles represent medians) and between model predictions and group means (right).

7 Experiment 3

In Experiments 1 and 2 deterministic causes were tested because we aimed to isolate the influence that the introduced temporal factors exert on singular causation judgments. However, causal power information should also play a crucial role in singular causation judgments according to our model. A central new prediction of our model is that causal power and temporal information are expected to interact. The goal of Experiment 3 was to explore this core property of the model.

A scenario was tested in which both causes either had a causal power of $w_C = w_A = 0.83$ or of $w_C = w_A = 0.5$. Causal latency was manipulated by using the first pair of gamma distributions (G1 vs. G5) shown in Fig. 9. The causal latency of the target cause was either associated with G1 or G5, while the causal latency of the alternative cause always followed the complementary distribution of the pair. This combination of causal powers and relative causal latencies of the potential causes leads to a predicted interaction effect that is shown in the middle panel of Fig. 12. The x-axis of Figure 12 shows the relative causal latency of the target cause, that is, whether it followed distribution G1 or G5. As can be seen, when the target cause’s relative causal latency is high (i.e., its causal latency follows G5 while the causal latency of the alternative cause follows G1), the generalized model predicts that ratings should be higher when the causal power of the target cause is low than when it is high (black vs.

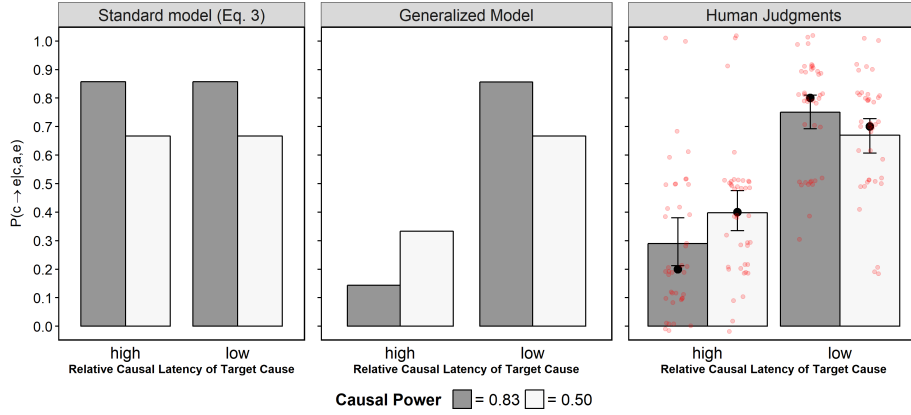


Figure 12. Model predictions and results (bars represent mean singular causation ratings; error bars denote 95% bootstrapped *CI*s; red points show jittered individual ratings; black points show medians) of Exp. 3. The x-axis shows the relative causal latency of the target cause.

grey bars in Fig. 12). Conceptually, this condition represents a scenario in which a reliably working target cause is competing with an alternative cause that not only is also working reliably but that also tends to operate quicker than the target cause. This should make it very unlikely that the target cause was the singular cause of the effect on an occasion on which both factors were present. Mathematically, the product that is subtracted from w_C in the high-power condition (black bar in the left pair) is sufficiently large so that the numerator ends up smaller ($0.83 - 0.83 \cdot 0.83 \cdot 0.99 = 0.15$) than in the low-power condition (grey bar in the left pair; $0.5 - 0.5 \cdot 0.5 \cdot 0.99 = 0.25$). Furthermore, the denominator is smaller in the low-power condition than in the high-power condition ($0.5 + 0.5 - 0.5 \cdot 0.5 = 0.75$ vs. $0.83 + 0.83 - 0.83 \cdot 0.83 = 0.97$), which means that the numerator in the low-power condition is increased more strongly than in the high-power condition. Fig. 12 also shows that the new model predicts the reversed effect when the relative causal latency of the target cause is low (i.e., when it follows G1; see right pair of bars in the middle panel of Fig. 12). Here, the product that is subtracted in the numerator from the target cause’s causal power becomes small because the target cause is operating quicker than its competitor. The causal-latency

advantage that the target cause is having allows it to manifest its high causal power. In this condition, our model makes the same predictions as the standard power PC model of causal attribution (Eq. 3): subjects should give higher singular causation ratings in the high causal power condition than in the low causal power condition. Finally, Fig. 12 shows that the new model also predicts a main effect of causal latency: Causes that on average tend to precede the efficacy of their competitors (right pair of bars) should receive higher singular causation ratings than causes whose competitors tend to preempt them (left pair of bars). Finally, a main effect of causal power is predicted by the standard but not by our generalized model.

7.1 Materials

7.1.1 Participants. One hundred and sixty subjects ($M_{age} = 38.14$, $SD_{age} = 12.20$, 116 female, 44 male) were recruited via Prolific (www.prolific.ac). This sample size ($n = 40$ subjects per condition) allows it to detect medium effects of $d = 0.60$ for simple group comparisons done with directed t-tests with more than 80 percent test-power. Subjects were at least 18 years old, had at least an A-level degree, were native English speakers, and had a work acceptance rate of at least 90 percent. Subjects were paid £ 1.20 for participation.

7.1.2 Design, Materials, and Procedure. Subjects were randomly assigned to one of four conditions ($n = 40$) that resulted from a 2 (causal power of the two causes: $w_C = w_A = 0.83$ vs. $w_C = w_A = 0.50$) \times 2 (relative causal latency of the target cause: high vs. low) between-subjects design. The following factors were counterbalanced between subjects: the order of presentation of the two causes in the learning phase (i.e., tower “North” first vs. tower “South” first); which cause had the lower/ higher relative causal latency; the target cause to which the main test question referred; the orientation of the rating scale belonging to the main test question (high confidence on the left vs. on the right side of the scale). This yielded 32 conditions in total.

The materials and procedure were largely the same as those of Experiments 1 and 2 with the following exceptions: First, information about the probabilistic causal nature

of the towers was added to the instructions. Participants read that pigeons might get lost on their way to the palace and that it would therefore be important to learn about the pigeons' arrival rates. Secondly, the learning task was modified so that causal power and causal latency information were conveyed together.

Other than in Experiment 2, subjects observed 24 pigeons per tower. This was done to ensure that all participants observe a sufficient number of successful pigeons to be able to learn the towers' causal latencies. Subjects in the high-power condition observed 20 successful pigeons per tower, whereas subjects in the low-power condition observed 12 successful pigeons per tower. The flight durations corresponded to 12 or 20 percentiles of the respective causal latency distributions. Whenever a pigeon failed to reach the palace, the words "pigeon probably lost" were displayed five seconds after the pigeon had been sent out. This duration corresponded to the 99.9th percentile of the slower distribution G5. The singular causation test questions were identical with the ones used in Experiments 1 and 2. Subjects were again asked to assess the probability that tower "North"/ "South" caused the alarm on a singular occasion on which both towers had sent their pigeons at the same time.

Additionally, on a separate screen subjects were asked to estimate the causal powers of the two causes because we wanted to control for the possibility that subjects' representations of the causal powers of the two causes might have been influenced by the causes' causal latencies. The new model considers causal power and causal latency to be two independent properties of causes: While causal power solely reflects causal strength (cf. Cheng, 1997), causal latency solely reflects the time it takes a cause to generate its effect. Yet, subjects might not have kept causal power and latency information apart, and might have, for instance, represented a relatively slow cause as less powerful than a relatively fast cause. This would be problematic for the generalized model as the results could then be explained by perceived differences in the causal powers alone. Empirical evidence for the possibility of an influence of delays on causal strength representations comes, for example, from the paper by Shanks et al. (1989), in which it was shown that the degree to which subjects attribute causality to a self-initiated action was correlated

negatively with the perceived delay between the action and the potential effect (see also Reed, 1992, 1999). Likewise, Lagnado and Speekenbrink (2010) showed that longer delays between the onsets of a potential cause and a potential effect lead to lower causation ratings, which in their paradigm was explained by possible hidden factors impinging on the causal process. Thus, the influence of temporal delays on causal strength judgments seems to be due to uncertainty about the underlying causal structure, especially about the possibility of unobserved factors.

To rule out uncertainty about the causal structure, we explicitly introduced the complete general causal structure in the instructions, ruling out causal influences by unobserved factors. Subjects in the present study were instructed that the two (observed) causes were the only causes of the effect. To assess subjects' causal power representations, they were asked the following question for each tower, with the order of presentation being randomized: "Based on what you have learned: how many out of 10 letter pigeons sent from Tower 'North' ('South') would make it to the palace?". Ratings were provided on an eleven-point scale (0 to 10).

7.2 Results and Discussion

The results are summarized in the right panel of Fig. 12. As can be seen, the singular causation ratings followed the predictions of the new model, whereas the standard power PC model of causal attribution does not explain the results. As for the predicted interaction, Fig. 12 shows that when the relative causal latency of the target cause was high, ratings in the low-causal power condition were higher ($M = 0.40$, $SD = 0.23$, $Md = 0.40$, $95\% CI = 0.07$) than those in the high-causal power condition ($M = 0.29$, $SD = 0.26$, $Md = 0.20$, $95\% CI = 0.08$). The reversed pattern was obtained when the relative causal latency of the target cause was low ($M = 0.67$, $SD = 0.20$, $Md = 0.70$, $95\% CI = 0.06$ when causal power was 0.50 vs. $M = 0.75$, $SD = 0.19$, $Md = 0.80$, $95\% CI = 0.06$ when causal power was 0.83). A planned contrast testing the predicted interaction was significant, $t(156) = 2.68$, $p < .01$, $d = .32$. However, Fig. 12 also shows that the simple effects qualifying the interaction

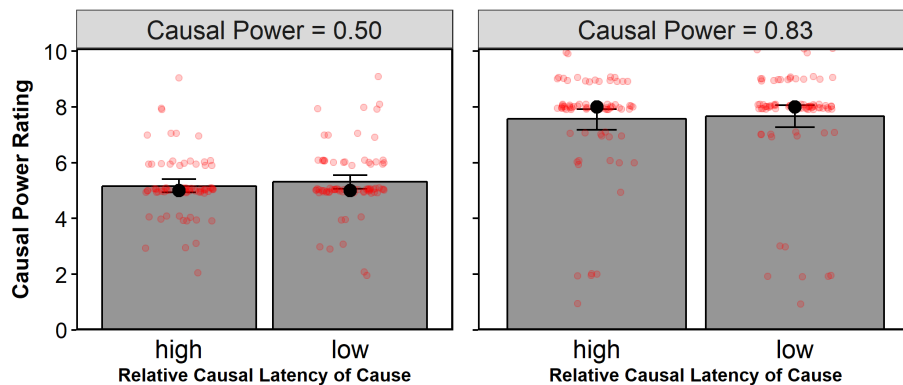


Figure 13. Results (bars represent means; error bars denote 95% bootstrapped *CI*s; red points show jittered individual ratings; black points show medians) for the causal-power question in Exp. 3. The left panel shows the ratings for the condition in which both causes had a causal power of 0.50. The right panel shows the ratings for the condition in which both causes had a causal power of 0.83. The x-axis shows the relative causal latency of the evaluated cause.

were smaller than predicted by the model ($t(156) = 2.17, p = .016$ one-sided, $d = 0.44$ for the condition the target cause had a relatively high causal latency vs. $t(156) = 1.62, p = .05$ one-sided, $d = 0.41$ for the condition in which the target cause's causal latency was relatively low).

Fig. 12 shows that the results also indicate a main effect of the relative causal latency of the target cause. A planned contrast testing this predicted main effect was significant, $t(156) = 10.47, p < .001, d = 1.62$. Finally, the main effect of the causal power of the causes, which is predicted only by the standard but not by the generalized model, was not found.

The causal power ratings that subjects made are summarized in Fig. 13. As can be seen, subjects' causal power representations were not distorted by the different causal latencies of the two causes. In the low-causal power condition (left panel in Fig. 13), the mean ratings for the cause with the higher latency were $M = 5.16$ ($SD = 1.07$). The mean ratings for the cause with the lower latency were $M = 5.31$ ($SD = 1.18$). In the high-causal power condition (right panel in Fig. 13), the mean ratings were $M = 7.58$

($SD = 1.75$) for the cause with the higher causal latency and $M = 7.66$ ($SD = 1.86$) for the cause with the lower causal latency. A mixed ANOVA with the relative causal latency of the target cause as within-subject factor and the causal powers of the causes as between-subject factor yielded a main effect only for the latter factor, $F(1, 158) = 122.68$, $p < .001$, $d = 1.6$, confirming that subjects in the high-causal power condition gave higher causal power ratings than subjects in the low-causal power condition. Fig. 13 also shows that subjects' causal power ratings were close to the normative values of 0.83 and 0.50. Importantly, there was no main effect of the relative causal latency of the target cause, $F(1, 158) = 1.38$, and also no interaction effect of the relative causal latency of the target cause factor and the causal power of the cause factor, $F(1, 158) < 1$. These results rule out that subjects' singular causation ratings can be explained by differences in the perceived causal powers of the causes.

8 Experiment 4

Although the pattern of singular causation judgments in Experiment 3 was captured well by the new model, the predicted interaction effect turned out to be relatively weak in the data. The predicted interaction follows from a core principal of the new model, however, which is why we decided to try to replicate the results in a fourth, pre-registered experiment that investigated a set of causal power parameters that should lead to a more substantial interaction effect. The same causal latency distributions were used as in the previous study but this time causes were contrasted that either had a power of 0.93 or 0.40. The model predictions are depicted in Figure 14. The experiment was pre-registered at the Open Science Framework (OSF; <https://osf.io/>). The pre-registration can be accessed under <https://doi.org/10.17605/OSF.IO/N93BU>.

At the end of Experiment 3 subjects were asked to estimate the causal powers of the two causes because it was important to control for the possibility that subjects' causal power representations might have been influenced by the differences in the causal latencies. The obtained results ruled out that subjects conflated the two types of

information when uncertainty about causal structure was reduced. However, it is possible that there is an influence in the other direction. Differences in the causal powers might have an impact on causal latency representations. This is not unlikely in the present paradigm because trials in which a cause failed to generate the effect ended after a duration that corresponded to the 99.9th percentile of the slower causal latency distribution (G5). Consequently, subjects in the low-causal power condition spent more time on the task than subjects in the high-causal power condition. The present study therefore assessed how strongly subjects' causal latency representations might be affected by causal power information.

8.1 Materials

8.1.1 Participants. Three hundred and eighty-four subjects ($M_{age} = 34.25$, $SD_{age} = 11.38$, 182 female, 200 male, two subjects did not indicate their gender) were recruited via Prolific (www.prolific.ac). Subjects were at least 18 years old, had at least an A-level degree, were native English speakers, and had a work acceptance rate of at least 90 percent. They were paid £ 1.20 for participation. The sample size calculation was done with the program G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) and based on the effect sizes found in a set of pilot studies conducted after the third experiment prior to the pre-registration of the new study. Both Experiment 3 and these pilot studies indicated that the smallest difference between the means is to be expected in the condition in which the target cause has a relatively low causal latency (see right pair of bars in Figure 12). With the calculated sample size, an independent-samples t-test (one sided) comparing the mean difference in this condition would detect a medium effect of $d = 0.5$ (alpha = 0.05) with about 95 percent probability.

8.1.2 Design, Materials, and Procedure. The study design was identical to the one of Experiment 3, that is, subjects were randomly assigned to one of four conditions ($n = 96$) that resulted from a 2 (causal power of the two causes: $w_C = w_A = 0.93$ vs. $w_C = w_A = 0.40$) \times 2 (relative causal latency of target cause: high vs. low) between-subjects design.

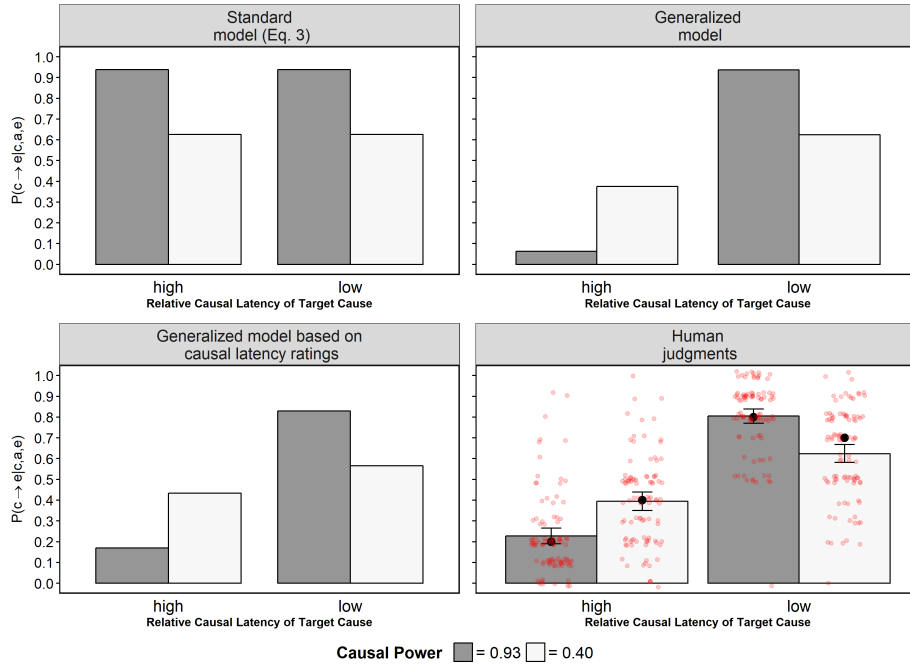


Figure 14. Model predictions and results (bars represent mean singular causation ratings; error bars denote 95% bootstrapped CIs ; red points show jittered individual ratings; black points show medians) of Experiment 4. The x-axis shows the relative causal latency of the target cause.

As this experiment contrasted the causal power parameters of 0.93 and 0.40, the learning phase differed from the one in Experiment 3. In the condition in which the causes had a causal power of 0.93, subjects learned that each cause was successful in bringing about the effect in 28 out of 30 attempts. In the low-causal power condition, by contrast, each cause generated the effect in only 12 out of 30 occasions.

After subjects had answered the singular causation query, which was the same as in the previous experiments, they were asked a final control question referring to the causal latencies of the two causes. Subjects were asked to imagine an occasion on which both towers simultaneously sent a pigeon, and then should indicate on an eleven-point scale with endpoints “definitely the pigeon from Tower ‘North’” and “definitely the pigeon from Tower ‘South’” (the midpoint was labeled “50:50”) which pigeon they think would arrive at the palace first. Answers to this question allowed us to estimate the α

parameter. The orientation of the scale (whether tower North was on the left or right side of the scale) was randomly varied between subjects.

8.2 Results and Discussion

The results are summarized in the fourth panel of Fig. 14 where it can be seen that the replication was successful. The results of a 2 (causal power of the two causes: 0.93 vs. 0.40) \times 2 (relative causal latency of target cause: high vs. low) factorial ANOVA confirmed that all effects predicted by the new model were significant. There was a main effect of “relative causal latency of target cause”, $F(1, 380) = 381.02$, $p < .001$, $d = 2.0$, confirming that singular causation ratings were overall higher when the target cause had a shorter causal latency than the alternative cause. Importantly, the interaction of “causal power” and “relative causal latency of target cause” was also significant, $F(1, 380) = 71.19$, $p < .001$, $d = .87$. Contrast analyses testing the simple effects that qualify the interaction showed that the difference between high and low causal power was significant in both “relative causal latency of target cause” conditions, $t(380) = 5.74$, $p < .001$. The effect size $d = 0.85$ for the condition in which the target cause’s causal latency was higher than the one of the alternative cause ($M = 0.39$, $SD = 0.22$, $Md = 0.40$, $95\% CI = 0.04$ when causal power was 0.40 vs. $M = 0.23$, $SD = 0.20$, $Md = 0.20$, $95\% CI = 0.04$ when causal power was 0.93), and $t(380) = 6.20$, $p < .001$, $d = 1.05$ for the condition in which the target cause’s causal latency was lower than the one of the alternative cause ($M = 0.62$, $SD = 0.22$, $Md = 0.70$, $95\% CI = 0.04$ when causal power was 0.40 vs. $M = 0.81$, $SD = 0.17$, $Md = 0.80$, $95\% CI = 0.03$ when causal power was 0.93). As predicted by the new model, there was no main effect of causal power, $F(1, 380) < 1$. Furthermore, it can also be seen that the effects were overall larger than the ones we observed in Experiment 3, which indicates that subjects were indeed sensitive to the causal power difference. In Experiment 3, the size of the interaction effect was $d = 0.32$, whereas it was $d = .87$ in the present study.

The ratings for the causal latency question that subjects made in the two different

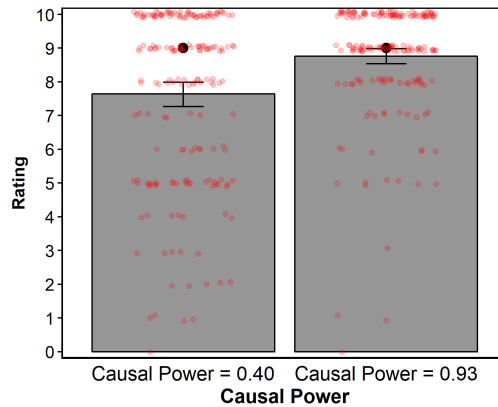


Figure 15. Results (bars represent means; error bars denote 95% bootstrapped *CI*s; red points show jittered individual ratings; black points show medians) for the causal-latency question that we asked in the end of Exp. 4. The x-axis shows the respective causal-power conditions. High ratings are associated with the cause with the lower causal latency.

causal power conditions are summarized in Fig. 15. The figure shows the results averaged over the different balancing factors. High ratings are associated with the cause with the lower causal latency (i.e., the one that operated quicker). First of all, it can be seen that ratings were high in both conditions, confirming that subject correctly identified the cause that had the smaller causal latency. Yet, it can also be seen that the causal power difference led to a small distortion. When the causal power of the two causes was high, subjects were more likely to give a rating associated with the cause with the lower causal latency ($M = 8.77$, $SD = 1.69$) than when the causal power of the causes was low ($M = 7.64$, $SD = 2.54$). Subjects' causal latency ratings were used to estimate the α parameter in our model (0.877 and $1 - 0.877$ for high vs. low alpha in the high-causal power case; 0.764 and $1 - 0.764$ for high vs. low alpha in the low-power case) to see how this would change the predictions for the singular causation ratings. The model predictions based on subjects' causal latency ratings are depicted in the third panel in Fig. 14. It can be seen that this model explains subjects' singular causation judgments even better than the model that relies on the objective parameter

values. The results of this experiment demonstrate again the validity of the new model.

9 General Discussion

The present paper argues for the view that assessing which of a set of potential cause factors actually caused a target effect on a singular occasion requires the integration of two different types of information: One relevant type of information is the causal strength of the potential causes: If an outcome can be explained by either of two potential causes, then the outcome is more likely due to the stronger cause. The standard causal power theory of attribution by Cheng and Novick (2005) provides a formal analysis of how the causal strengths of the potential causes should be weighed against each other to reach a decision. Considering only causal strength is not enough, though. The second crucial type of information that needs to be incorporated is information about temporal relations. Temporal information is relevant because it would otherwise be impossible to distinguish between situations in which a sufficiently strong cause actually succeeds in bringing about the effect and situations in which a likewise sufficient competitor managed to preempt the target cause. We here proposed a formal model that incorporates the probability of preemption to estimate how likely it is that the target cause actually caused the target outcome. We showed that the probability of preemption can be obtained by the combination of causal strength along with temporal information about the potential causes' onset times and their causal latencies. The results of the reported experiments demonstrate that the new model makes accurate predictions of subjects' singular causation judgments. Interestingly, the model was accurate even in the computationally quite demanding situations where all its parameters were probabilistic (Experiments 3 and 4).

9.1 Limitations and avenues for future studies

9.1.1 Modeling other types of situations. The type of test situations we investigated represents only a subset of possible situations. For example, one noteworthy characteristic of our test scenarios was that all potential causes of the effect were actually observed. In most real-world situations, however, reasoners observe only a

subset of the potential causes and thus are typically confronted with uncertainty concerning the presence of further alternative causes. Although we have not considered situations with unobserved background causes here, the new model can be applied to such situations, too. In situations in which only the target cause is observed while alternative causes remain unobserved, information about the temporal distribution of the effect in the absence of the target cause needs to be considered to estimate α . As Bramley et al. (2018) have shown in their recent article on general causal structure induction, the temporal distribution of the effect in the target cause's absence can be modeled with exponential functions. The observed causal-delay distribution in the presence of the target cause would then be a mixture of the particular gamma distribution of the target cause's causal latency and the exponential background distribution. The estimation of the α parameter would require as a first step the decomposition of the mixed distribution into its elements. Thereafter, the same algorithm that we used to estimate α in the present studies can be applied. We plan to test such situations in future experiments and expect similar findings: target causes with a small causal latency should more likely be viewed as singular causes than otherwise identical causes with higher causal latency, as the latter are more likely to be intercepted by unobserved background causes (see Lagnado & Speekenbrink, 2010).

Furthermore, the present studies tested only static target situations, as subjects were asked to make singular causation judgments based on previously acquired knowledge about strength, delays and onsets. The structure of the target situation thus resembled a classical crime scenario, in which the detective enters the scenery and tries to identify the perpetrator after all events had unfolded. We focused on this type of test situation because it is the standard situation analyzed by the standard causal power model of attribution (Cheng & Novick, 2005) and because it is the type of situation that was tested in previous related studies (e.g., Holyoak et al., 2010, or also Meder et al., 2014, who tested diagnostic inferences). In many real-life situations, however, singular causation queries arise in dynamic contexts. A crucial feature of dynamic test cases is that the onsets and delays are experienced online prior to responding to a

singular causation query. This way the interaction between prior beliefs about causal strength and temporal parameters could be compared to the actually observed sequence of events. Imagine a target cause with an average causal strength of $w_C = 0.90$ and an expected causal delay of 20 minutes (e.g., a dose of ibuprofen). Now subjects observe that the relief from pain occurs after two hours, however. The unexpected long delay may suggest that the causal strength of the drug was zero on this occasion, and therefore should weaken the belief that the drug was the singular cause of the effect.

9.1.2 The role of causal mechanisms. Another question that we did not address in the present paper concerns the role of causal mechanism information in the assessment of singular causation. After all, causal mechanism information has been identified as an important type of information in causal reasoning (see Johnson & Ahn, 2017, for an overview). Different studies have shown that reasoners regard mechanism information as particularly relevant when making causal attributions (Ahn, Kalish, Medin, & Gelman, 1995; Johnson & Keil, 2018). Likewise, in a recent article the philosopher Nancy Cartwright (2017) has listed mechanistic information among the factors that help assess singular causation.

We think that the influence of mechanism information can be modeled by the addition of variables to our model. Causal mechanisms can be represented by additional variables interpolated between the cause and the effect factors, thereby turning a previously direct causal connection (e.g., $C \rightarrow E \leftarrow A$) into an indirect one ($C \rightarrow M_C \rightarrow E \leftarrow M_A \leftarrow A$). Consider our standard scenario in which a reasoner knows that C , A , and E occurred on a singular occasion and they are asked to respond to the query how confident they are that c instead of a caused e on this occasion. Knowledge about the status of the mechanism variable M_A connecting A and E would be helpful here because knowing the status of M_A constrains a 's actual causal power on the target occasion. If the causal chain $A \rightarrow M_A \rightarrow E$ honors the causal Markov condition and if there is only one "mechanism path" leading from A to E (on which M_A lies), the absence of M_A ($\neg m_A$) implies that a had a causal power of zero in this case and can therefore be ruled out as a singular cause of e . Consequently, whenever c is the

only remaining potential cause of the effect, one can safely conclude that e was caused by c . This normative reasoning strategy has been called “eliminative abduction” or “Holmesian inference” in the literature (Bird, 2010) and our model is in accordance with it. To see this, consider how Eq. 5 would have to be modified to capture such a situation. What would change is that another conditional element would have to be added to $P(c \rightarrow e|c, a, e)$ turning this expression into $P(c \rightarrow e|c, a, \neg m_A, e)$. On the right-hand side of the equation, the unconditional causal power of A , w_A , would have to be substituted by A ’s power conditional on the absence of M_A , $w_{A|\neg M_A}$. Assuming that the causal Markov condition holds and that there is only a single mechanism path connecting A and E , $w_{A|\neg M_A}$ would amount to zero and $P(c \rightarrow e|c, a, \neg m_A, e)$ would reduce to $\frac{w_C}{w_C} = 1$. Of course, even in cases in which A produces E via multiple mechanism paths, assessing the status of M_A can be helpful. If a situation can be conditionalized on the absence of M_A , this should increase $P(c \rightarrow e|c, a, \neg m_A, e)$, as the causal Markov condition implies that $w_{A|\neg M_A} < w_A$.

Causal mechanism information might not only constrain the causal strength parameters in the model but also the parameter α . The reason for this is that a causal mechanism connecting a target cause and a target effect might involve multiple causal paths which might differ not only with respect to their causal powers but also with respect to their causal latencies. Take the example of a coroner who wants to find out whether a victim displaying an abdominal gunshot wound actually died because of the gunshot. The causal strength with which bullets kill their victims is surely quite high on average, but also fairly variable. Similarly, bullets typically bring death to their victims relatively quickly, but not always. Depending on the organs that are damaged, both the probability of dying from a gunshot and the latency can be high or low. A case in which the bullet took a straight path to the victim’s heart will probably lead to higher confidence in a singular causal link between gunshot and death than a case in which the bullet left all organs intact. Bullets that stop their victims’ hearts are not only more effective, they also bring death so quickly that there is less room for alternative causes to strike (e.g., an assassin might first have administered a lethal dose

of poison but then decided to shoot the victim). Manipulating and testing the influence of causal mechanism information on singular causation judgments will be an interesting avenue for future studies.

9.1.3 Singular events and their corresponding types. The account we are presenting here commits to the philosophical point of view that singular causation is not directly observable, and that the corresponding judgments therefore need to involve general-level causal knowledge gained from multiple observations (cf. Danks, 2017). Here the problem arises that a suitable type-level causal variable or reference class need to be chosen. However, every singular case can be associated with infinitely many type-level variables. In the case of Peter's smoking as a potential singular cause of his lung cancer, Peter might be classified as an instantiation of the type smoker or more specifically as an instantiation of the type smoker who smokes more than ten cigarettes but less than a whole pack per day. An important problem here is that causal strength estimates in most cases vary depending on the chosen reference class (see Cartwright, 1989). In the present research we have not studied the problem of how reasoners select reference classes. This is an interesting question for future studies, however. A plausible hypothesis one could test is that people will prefer to associate singular cases with a reference class that is maximally homogeneous. Although this hypothesis is a plausible first pass, it surely needs qualification. For example, narrowing the reference class too much so that the target case represents its only member would also not work as it would render the reference class useless.

9.2 Conclusion

Assessing whether two singular events c and e were actually causally connected or whether their co-occurrence is a coincidence requires the application of general causal background knowledge about the strength with which the potential cause C tends to generate the effect relative to possible alternative cause A . This is not enough however to respond to a singular causation query because not even deterministic causes are necessarily causal on a singular occasion. A simultaneous sufficient alternative cause

might have preempted the target cause in its efficacy. To account for this possibility, causal strength information needs to be combined with temporal information about the potential causes. Two relevant types of temporal information are onset differences between the potential causes and causal latencies. Our new generalized model of singular causation judgments combines causal power and temporal information to compute the probability with which a target cause was preempted by an alternative cause. Four experiments demonstrated that reasoners integrate and apply causal strength and temporal knowledge as predicted by our new model.

References

- Ahn, W.-k., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.
- Bird, A. (2010). Eliminative abduction: examples from medicine. *Studies in History and Philosophy of Science*, *41*, 345–352.
- Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). The role of time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *44*, 1880–1910.
- Buehner, M. J. (2017). Space, time, and causality. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 549 – 564). New York: Oxford University Press.
- Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*, 353–378.
- Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Clarendon Press.
- Cartwright, N. (2017). Single case causes: What is evidence and why. In H. K. Chao & J. Reiss (Eds.), *Philosophy of science in practice*. (pp. 11–24). Cham: Springer.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.
- Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*, 694–706.
- Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 201–215). New York: Oxford University Press.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
- Glymour, C. (2003). Learning, prediction and causal bayes nets. *Trends in cognitive sciences*, 7, 43–48.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 334–384.
- Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, 30, 1128–1137.
- Halpern, J. Y. (2016). *Actual causality*. Cambridge, MA: MIT Press.
- Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66, 413–457.
- Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. Oxford: Oxford University Press.
- Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98, 273–299.
- 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, 702–727.
- Johnson, S. G., & Ahn, W.-k. (2017). Causal mechanisms. In M. R. Waldmann (Ed.), (pp. 127–146). New York: Oxford University Press.
- Johnson, S. G., & Keil, F. C. (2018). Statistical and mechanistic information in evaluating causal claims. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 618–623). Austin, TX: Cognitive Science Society.
- Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 565–602). New York: Oxford University Press.

- Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*, 184–195.
- Lewis, D. (1973). Causation. *The journal of philosophy*, *17*, 556–567.
- Liljeholm, M., & Cheng, P. W. (2007). When is a cause the “same”? coherent generalization across contexts. *Psychological Science*, *18*, 1014–1021.
- Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 415–432). New York: Oxford University Press.
- Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*, 277–301.
- Paul, L. A., & Hall, E. J. (2013). *Causation: A user’s guide*. Oxford University Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge, NY: Cambridge University Press.
- Reed, P. (1992). Effect of a signalled delay between an action and outcome on human judgement of causality. *The Quarterly Journal of Experimental Psychology Section B*, *44*, 81–100.
- Reed, P. (1999). Role of a stimulus filling an action-outcome delay in human judgments of causal effectiveness. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*, 92.
- Schaffer, J. (2003). Overdetermining causes. *Philosophical Studies*, *114*, 23–45.
- Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*, 139–159.
- Shortle, J. F., Thompson, J. M., Gross, D., & Harris, C. M. (2018). *Fundamentals of queueing theory*. John Wiley & Sons.
- Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

- Sloman, S., & Lagnado, D. A. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223–247.
- Spellman, B. A., & Kincannon, A. (2001). The relation between counterfactual (but for) and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems: Causation in Law and Science*, *64*, 241–264.
- Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*, 242–257.
- Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.
- Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

Appendix D

Stephan and Waldmann (2016)

Stephan, S., & Waldmann, M. R. (2017). Answering causal queries about singular cases. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2795–2801). Austin, TX: Cognitive Science Society.

Answering Causal Queries about Singular Cases

Simon Stephan (simon.stephan@psych.uni-goettingen.de)

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)

Department of Psychology, University of Göttingen,

Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Queries about singular causation face two problems: It needs to be decided whether the two observed events are instantiations of a generic cause-effect relation. Second, causation needs to be distinguished from coincidence. We propose a computational model that addresses both questions. It accesses generic causal knowledge either on the individual or the group level. Moreover, the model considers the possibility of a coincidence by adopting Cheng and Novick's (2005) power PC measure of causal responsibility. This measure delivers the conditional probability that a cause is causally responsible for an effect given that both events have occurred. To take uncertainty about both the causal structure and the parameters into account we embedded the causal responsibility measure within the structure induction (SI) model developed by Meder et al. (2014). We report the results of three experiments that show that the SI model better captures the data than the power PC model.

Keywords: causal inference, generic causation, singular causation, actual causation, causal responsibility, causal attribution, Bayesian modeling

Imagine that you wake up one morning and recognize that you are haunted by a mean twinge in your head. You also know that you drank too many glasses of wine last night. Now the question arises whether your behavior last evening is causally responsible for your headache this morning. This causal query targets a *singular* instance in which one event at a specific spatio-temporal location may have caused another event that followed. The general problem with singular causal queries is that a co-occurrence of the particular events by itself does not guarantee causation. It may just be a *coincidence*. How confident can you be that this singular case of having drunk wine is the cause of your headache?

One important source that should influence our confidence is past knowledge about the contingency between drinking wine and headache. This is *generic* causal knowledge. It could either refer to cases of presence and absence of drinking and headache in an observed sample of people, or, even better with respect to the example given above, to a sample of these events in your life. The contingency in your life provides the best estimate for a generic causal relation between drinking and headache in your body.

However, knowing that there is a generic causal relation does not necessarily imply that a singular co-occurrence of the target events is causal. Unless the generic relation is deterministic, the co-occurrence may still be a coincidence. Thus, our causal judgment about singular causation must take this possibility into account.

Psychological research on causal inference adheres to two different theoretical frameworks to characterize the reasoning processes (see Waldmann, in press; Waldmann & Hagmayer,

2013; Waldmann & Mayrhofer, in press). One approach assumes that causal knowledge is grounded in knowledge about causal dependencies gleaned from observed contingencies. According to this approach, causes are *difference makers* that raise or lower the probability of an effect (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Meder, Mayrhofer, & Waldmann, 2014).

A different approach assumes that causal knowledge is based on a search for mechanisms and processes linking causes and effects (e.g., Ahn, Kalish, Medin, & Gelman, 1995). The two approaches need not be incompatible. Often mechanism knowledge is based on generic information about causal chains. Specific mechanisms are then simply instantiations of generic chain knowledge.

Mechanism knowledge is often not available. But even in cases in which information about intervening variables of a causal chain is in fact accessible, the question arises again how we should distinguish causation from coincidence. The joint occurrence of all elements of a chain certainly makes the possibility of a coincidence extremely unlikely, but it is in principle still a possibility. Thus, the general problem of how we should apply generic knowledge to singular cases still needs to be addressed, regardless of whether we solve this problem for a direct causal link or a causal chain.

In the present paper, we investigated how people exploit generic causal information when estimating singular causation. We here focus on cases for which it is known that both the cause event and the effect event are present, but the question needs to be answered how much confidence we can have that the co-occurrence is due to causation and not a coincidence.

Cheng and Novick (2005) have proposed a measure of *causal responsibility* that helps to answer this question. The measure is based on the assumptions of power PC theory (Cheng, 1997). It delivers the conditional probability that a cause event c is causally responsible for an effect event e given that a reasoner knows that both have occurred, $P(c \rightarrow e|c, e)$. We think that this quantity underlies judgments of singular causation when only generic information is available. One shortcoming of this measure is that it does not incorporate the reasoners' *uncertainty* concerning the underlying causal structure and the corresponding parameters (see Griffiths & Tenenbaum, 2005). Therefore, Holyoak, Lee, and Lu (2010) have proposed a causal responsibility measure that also takes parameter uncertainty into account. We go one step further here, and test a model of causal responsibility that is sensitive to both parameter uncertainty and uncertainty about the existence of the causal link between factors C and E . This

model is based on the structure induction (SI) model of diagnostic reasoning developed by Meder et al. (2014). We present three experiments in which we empirically tested the power PC theory of causal responsibility against our SI model of singular causation.

The Power PC Model of Causal Responsibility

Causal power in Cheng’s (1997) theory can be understood as the probability that the target cause brings about the effect in the absence of all alternative causes of the effect. Fig. 1 (right) shows the basic causal model assumed by the power PC theory. The default assumption is that the target effect E can be independently produced by either the observed cause C with causal power (or strength) w_c or by alternative unobserved causes A with the power (or strength) w_a . It is further assumed that C and A occur independently and do not interact. Thus, C and A combine through a *noisy-OR* gate (see Griffiths & Tenenbaum, 2005; Pearl, 1988).

Based on the power PC framework, Cheng and Novick (2005) developed a measure of causal responsibility. Formally, causal responsibility is the proportion of occurrences of the effect due to the target cause C . Cheng and Novick (2005) presented formalizations of different kinds of causal responsibility. Here we are interested in how to answer the question whether an observed event c was causally responsible for an observed event e in cases in which both c and e were observed. Under the default assumptions of power PC theory, Cheng and Novick (2005) showed that this quantity is given by

$$P(c \rightarrow e|c, e) = \frac{P(c) \cdot w_c}{P(c, e)} = \frac{P(c) \cdot w_c}{P(c) \cdot P(e|c)} = \frac{w_c}{P(e|c)}, \quad (1)$$

where $P(c)$ equals b_c in Fig. 1, which denotes the base rate of C , and $P(c, e)$ denotes the joint probability of cause and effect. Since $P(c, e)$ can be rewritten as the product of $P(c)$ and the predictive probability $P(e|c)$, causal responsibility given the joint occurrence of c and e equals the causal power of c divided by the predictive probability of e given c .

The power PC theory only considers the model S_1 depicted in Fig. 1 (right) and uses maximum likelihood estimates for the parameters. It does not take into account uncertainty about the size of the parameters, nor about the causal structure (see Griffiths & Tenenbaum, 2005). Thus, it does not consider the possibility, depicted as S_0 in Fig. 1, that there is no causal arrow from C to E , meaning that all co-occurrences are due to coincidence.

A consequence of maximum likelihood estimates is that the theory predicts maximal values of causal responsibility for any contingency where the target cause is necessary (i.e., a table with an empty C cell) no matter how rarely co-occurrences of cause and effect are. Consider the two examples of such cases depicted in Fig. 2. In the left table, the effect has only occurred four times when the cause event was present. In contrast, in the right table the effect occurred sixteen times when the cause event was present. By applying the

equations, one can see that the prediction of maximal responsibility is a consequence of w_c being equal to $P(e|c)$ for any table with an empty C cell.

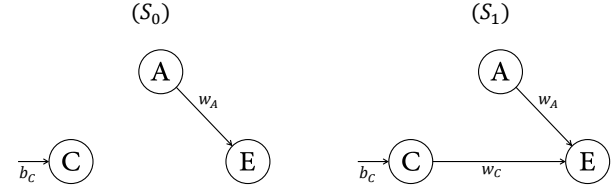


Figure 1: The two causal structures in the structure induction model. In S_0 , no causal relationship exists between C and E . In S_1 , C and E are causally connected. Node A represents unobservable background causes of E . The parameter b_c denotes the base rate of the cause, and w_c and w_a the causal powers of C and A , respectively.

The Structure Induction Model of Singular Causation

Contrary to the power PC theory, the structure induction (SI) model assumes that reasoners might be uncertain about both the underlying causal model and the parameters (Meder et al., 2014). Originally, the model was developed to model diagnostic inferences. Here, instead, we use the framework to model judgments about singular causation. Our key claim is that subjects assess singular causation by estimating causal responsibility within the SI model.

Causal queries about simple causal models with a single cause and a single effect are modelled by the SI theory as a Bayesian inference problem over the two mutually exclusive causal structures shown in Fig. 1 (see also Griffiths & Tenenbaum, 2005). The model formalizes the assumption that reasoners are uncertain about which of the two models underlies the data; they are also uncertain about the size of the parameters. We will briefly summarize the different computational steps of the model.

Parameter Estimation

In light of the possibility that either S_0 or S_1 might underlie the data, the model estimates the base rate and power parameters b_c , w_c , and w_a , separately for each causal structure (see Fig. 1). To express uncertainty, distributions of the parameters, rather than maximum likelihood point estimates, are inferred. According to Bayes’ rule, the posterior probability distributions for the parameters of each model given the data, $P(w|D)$, is proportional to the likelihood of the data given the set of parameters w , weighted by the prior probability of the structure:

$$P(w|D) \propto P(D|w) \cdot P(w) \quad (2)$$

$P(D|w)$ denotes the likelihood of the data given the parameter values for b_c , w_c , and w_a . $P(w)$ represents the prior joint

probability of the parameters which we set to flat, uninformative $Beta(1, 1)$ distributions.

Structure Estimation

The SI model separately derives the posterior probabilities for each causal structure. Applying Bayes' rule, the posterior probability for a causal structure is proportional to the likelihood of the data given the causal structure, weighted by the structure's prior probability:

$$P(S_i|D) \propto P(D|S_i) \cdot P(S_i) \quad (3)$$

$P(D|S_i)$ denotes the likelihood of the data given a particular structure, which is the integral over the likelihood function of the parameter values under the particular structure. $P(S_i)$ represents the prior probability of the structures. The model initially assumes that both structures have equal priors, i.e., $P(S_i) = 1/2$. When data are available, the posterior for a causal structure varies systematically with the observed contingency: the higher the contingency, the more likely S_1 becomes.

Causal Responsibility for Each Structure

Having estimated the parameters and the posteriors of the structures, the model computes causal responsibility separately for each parametrized structure using Equation 1. Under a noisy OR-parametrization Equation 1 can be rewritten as

$$P(c \rightarrow e|c, e) = \frac{w_c}{P(e|c)} = \frac{w_c}{w_c + w_a - w_c \cdot w_a}. \quad (4)$$

According to S_0 , there is no causal connection between C and E , and any co-occurrences of c and e are coincidences. Hence, $P(c \rightarrow e|c, e) = 0$ for S_0 . For S_1 , Equation 4 is applied. The estimation of causal responsibility is then derived by integrating over the parameter values.

Deriving a Single Value

The final output of the model is a single estimate of causal responsibility through integrating out the causal structures by summing over the derived values of $P(c \rightarrow e|c, e; S_i)$ for each structure weighted by each structure's posterior probability:

$$P(c \rightarrow e|c, e; D) = \sum_i P(c \rightarrow e|c, e; S_i) \cdot P(S_i|D). \quad (5)$$

The incorporation of structure and parameter uncertainty by the SI model leads to systematic deviations from the predictions of the power PC model (see Fig. 2). For an illustration, consider the left contingency table depicted in (a). Although the effect never occurred in the absence of the cause, only four co-occurrences are observed. This relatively low frequency of co-occurrence is the reason for the relatively high probability of S_0 . Hence, based on this information, it seems reasonable to be cautious about the causal relation of a singular co-occurrence. By contrast, consider the right table

in (a). Here again, the effect never occurred in the absence of C but it occurred frequently in its presence. In this case, the data suggest a strong generic causal relationship between the factors. Now, when presented with a singular case, relatively high confidence in a case of actual causation is to be expected.

Experiment 1

The experiment tests the SI model against the power PC theory as an account of how subjects answer singular causal queries. Fig. 2 shows the predictions of the models for the two data sets that we used. The power PC model predicts invariance whereas our SI model of singular causation predicts that judgments should vary. This is because the likelihood that the observed data pattern was produced by structure S_1 varied between the two different contingencies. We also wanted to address a second question: Do subjects differentiate between generic and singular queries? To test this, we compared a generic with a singular query in the experiment.

Design, Materials, and Procedure

Sixty-one subjects completed this online experiment for monetary compensation. Fifty-six subjects (mean age = 29.16 years, $SD = 8.4$ years, 28 were female) passed our embedded attention check and were included in the analyses. The contingencies (see Fig. 2) were varied between subjects.

We used a fictitious story of a single sea devil living in an aquarium. Our subjects were asked to assume they were biologists being interested in whether noise causes the fish's antenna to flash. Subjects read they should imagine having run a single-case study with one fish to test this hypothesis. We then instructed our participants that they will see the results of this study. They read that the fish was placed in front of a loudspeaker twenty times and that the loudspeaker was off at first and activated subsequently to assess the fish's reaction.

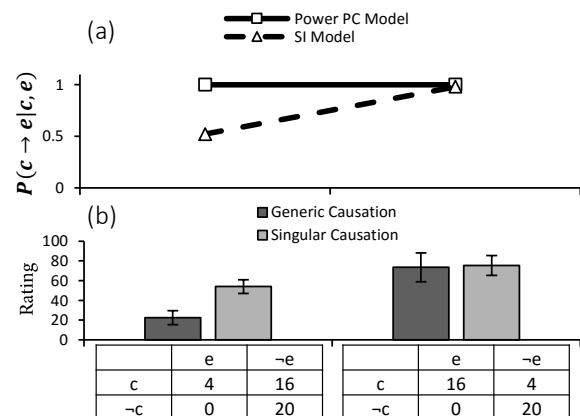


Figure 2: Predictions of the power PC and the SI model of singular causation are depicted in (a), the results in (b). Dark and light bars show means (95% CI) of the generic and singular causation ratings, respectively.

Having read the instruction, subjects saw the results of the observations arranged in a table with four columns and five rows. Below each depiction of the fish, located in front of a tiny loudspeaker, a little yellow scrip was placed that indicated the respective trial number. The initial screen showed the state of the fish as the loudspeaker was off ($\neg c$). Next, we presented the results as the loudspeaker was activated (c), symbolized by tiny sound waves. A yellow color of the fish's flash bulb indicated the flashing of the antenna.

To test intuitions about generic causation we asked subjects how appropriate it is to say (on a rating scale from 0 to 100) that noise is a cause of the flashing of the sea devil's antenna. The corresponding question targeting singular causation asked subjects to focus on the first trial of the observations in which the fish's antenna had flashed upon noise exposure. Subjects were requested to indicate (again using a scale ranging from 0 to 100) how appropriate it is to say that the noise had caused the flashing of the fish's antenna in this particular trial. Additionally, we also asked subjects about a trial in which the antenna did not flash upon noise exposure. Furthermore, we asked them to make a predictive judgment concerning a hypothetical new trial. Here, we asked how likely it is that the fish would flash its antenna again.

Results and Discussion

The results can be seen in Table 1 and Fig. 2. Fig. 2 shows that participants judged the generic-level causal relationship between noise and antenna flashing differently for the two contingencies. On average, ratings were higher in the high contingency condition compared to the low contingency condition. As predicted by the SI model, the singular cause ratings were also different in the two conditions.

A 2 (contingency) \times 4 (type of rating) mixed ANOVA with the second factor being varied within subject yielded a significant main effect of contingency, $F(1, 54) = 82.00, p < .001$, a significant main effect of rating, $F(3, 162) = 53.13, p < .001$, and also a significant interaction between contingency \times rating, $F(3, 162) = 15.74, p < .001$. A planned contrast comparing the singular causation ratings was also significant, $t(54) = 2.55, p = .01$, confirming that the ratings were higher in the high contingency condition. This difference is not predicted by the power PC model but it is predicted by the SI model.

An interaction contrast comparing the difference between the generic and the singular ratings was significant,

Table 1. Mean ratings (*SE* of the mean) obtained for the different questions in the experiment.

	condition	
	contingency: low	contingency: high
generic causation	22.40 (3.43)	73.55 (3.43)
singular causation (e, c)	54.00 (7.12)	75.48 (4.95)
singular causation ($e, \neg c$)	10.40 (3.58)	19.03 (5.36)
predictive probability	20.40 (3.13)	76.13 (2.48)

$t(54) = 3.60, p < .001$. As Fig. 2 shows, this finding is due to the fact that the mean rating obtained for the singular causation question was higher than the generic causation rating in the low contingency condition. While this indicates that participants indeed conceptualized the two causal queries differently it seems that subjects answered the generic question with a causal strength rather than a structure estimate. This may also explain why there was no difference between generic and the singular ratings in the high contingency condition. As can be seen in Table 1, the additional ratings for the predictive probability were in line with the empirical predictive probability obtained in the contingency tables. Interestingly, singular ratings differed from zero in both groups for the trial in which the fish did not show a flashed antenna after noise exposure. These ratings were also higher in the high contingency condition than in the low contingency condition. This seems to reflect an unpredicted influence of generic knowledge, but it should be noted that the ratings were the lowest in the set.

In sum, the results favor our SI model of singular causation. We obtained the predicted slope of the ratings that contradicts the power PC model predictions. Furthermore, the results obtained in the low contingency condition show that subjects differentiated between a generic and a singular causal query.

Experiment 2

In Experiment 1 we used data sets in which the cause appeared necessary for the effect. To test our model for data sets in which all event types exist, we used in the present experiment the three contingency tables shown in Fig. 3. Again, the power PC model predicts invariance. This time it predicts values of 0.83, whereas the SI model predicts a slope. For this experiment, we also wanted to compare the SI model with a Bayesian variant of the power PC model (Holyoak et al., 2010) that takes into account parameter uncertainty but not structure uncertainty. For this model, we also used uninformative priors over the parameters. As can be seen, the predictions of this model are very similar to the predictions of the power PC model but slightly lower on average.

A second change was that the generic information did not refer to a time series of trials involving a single individual, but a sample of different individuals. Thus, we were interested in testing how subjects use generic information about a sample of different fish to respond to a causal query about a singular case.

Design, Materials, and Procedure

Thirty-eight participants (mean age = 31.23 years, $SD = 8.90$ years, 14 were female) completed this online experiment. The data sets were presented to each subject.

In this experiment, all forty observations were shown in random order on the screen in a table with five columns and eight rows. Subjects should assume they were biologists who had studied the influence of chemicals on gene mutations in populations of sea devils. They were instructed that they will see the results of these studies. Subjects also read that the

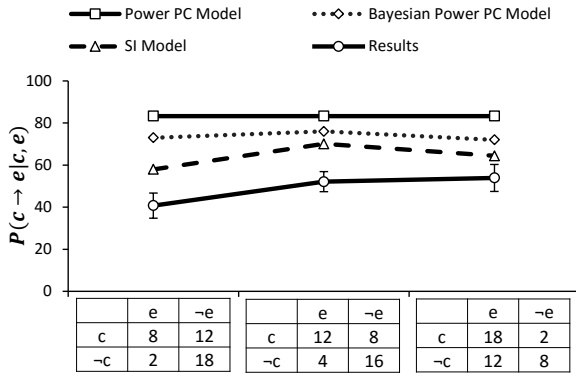


Figure 3: Model predictions and results (95% CI) for the different data sets used in Experiment 2.

chemical had been injected in half of the sample in each study and that the whole sample had later been tested for mutation. We used differently colored circles (red vs. blue) to indicate the presence vs. absence of gene mutations; a black margin around circles indicated chemical treatment. Finally, we instructed our participants that their task will be to provide a judgment concerning a single case.

For each data set, participants were asked to imagine a new individual fish which had ingested the chemical and also had the mutation. This time, we asked for the probability (on a scale from 0 to 100) that the chemical caused the gene mutation in this single case.

Results and Discussion

As can be seen in Fig. 3, ratings differed across the three data sets. A repeated-measure ANOVA with “data set” as within-subject factor was significant, $F(2, 74) = 4.19, p = .02$. This main effect was due to the ratings for the first data set being different from those for the second and the third data set, $t(37) = 2.77, p < .01$. The ratings for the second and the third data set did not differ, $t(37) = 0.39$. The dampening of the upward trend across the three data sets is consistent with our SI model, although the model predicts a slight downward trend between the second and third data set. In contrast to what we observed in the first experiment, the singular causation ratings for all three data sets were lower than predicted by the SI model. One explanation for this finding may be that the type of data presentation that we used made it hard for participants to grasp the contingencies precisely, because all observations were presented on the screen in random order and also in combination with abstract symbols.

In sum, the results again favor our SI model of singular causation over the power PC model and also over a Bayesian variant of the power PC model as an account of how subjects respond to causal queries about singular causation. The present experiment further demonstrates that not only time series data but also data about samples of different individuals are used to derive a prediction for a singular case.

Experiment 3

In the first two experiments we dissociated the SI model of singular causation from the power PC model by using data sets for which the power PC model predicts invariance. To obtain additional evidence for the validity of the SI model we conducted an experiment with data sets for which the SI model predicts invariance but the power PC model does not. We obtained invariant predictions of the SI model by counteracting a slight decrease in contingency across the two contrasted data sets (see Fig. 4) with an increase of sample size.

Apart from singular causal inferences, we also tested whether participants would infer a higher probability for the existence of a generic causal relationship (i.e., for S_1) for the larger sample (Fig. 4, right), which is predicted by the SI model for generic queries. Our model predicts an interaction effect between the type of question and data set.

Design, Materials, and Procedure

All subjects saw both data sets. The type of question (confidence in generic vs. singular causation) was manipulated between subjects. 101 subjects participated in the online study (mean age = 31.40 years, $SD = 11.10$ years, 43 were female) and provided valid data.

Participants read that they should imagine to be biologists who tested the influence of the chemical “Acrinazyl” on the expression of the gene ASPM in mice in an experiment using a sample of twenty test animals. They were instructed that one half of the sample served as a control group. We presented the two halves of the sample separately on the same screen. Different colors indicated the status of ASPM. In the generic condition, we asked subjects to indicate (on a slider ranging from 0 to 100) how confident they are that the chemical does indeed raise the probability of ASPM expression. We were hoping that this test question would less ambiguously refer to generic causal relations than the one we used in Experiment 1. In the singular condition, subjects were instructed that they should imagine having picked out a single mouse from the Acrinazyl group with ASPM being expressed. They were asked how much confidence they had that it was indeed

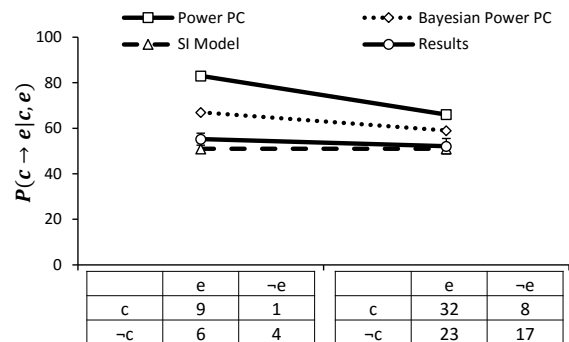


Figure 4: Model predictions and results (95% CI) for the different data sets used in Experiment 3.

the chemical that caused the gene expression in this particular mouse. Subjects indicated their confidence using a slider ranging from 0 to 100.

After participants had provided these ratings they read that they had conducted a second experiment with a larger sample size. We then showed them the results of the second study and asked them to re-assess their ratings in light of the results of this second study.

Results and Discussion

Fig. 4 shows that the singular ratings are well captured by the SI model. Moreover, as predicted by the SI model, the mean ratings for the generic causal question increased between data sets ($M = 63.54$, $SE = 1.00$; $M = 68.63$, $SE = 1.00$). A mixed ANOVA confirmed our predictions with a significant main effect for causal question, $F(1, 99) = 7.42$, $p < .01$, and a significant interaction between question type and data set, $F(1, 99) = 6.60$, $p = .02$. Furthermore, there was no difference between the singular causal ratings, $t(46) = 1.34$.

Overall, Experiment 3 was in line with the predictions of our SI model of singular causation. We also demonstrated again that subjects treat singular and generic causation queries differently.

General Discussion

We tested a new model of how people respond to queries about singular causation. Simply observing two consecutive events at a specific space-time location does not suffice. The question needs to be answered whether this co-occurrence manifests a causal relation or merely a coincidence. We argued that one relevant source of knowledge are generic causal relations. However, knowing, for example, that smoking generally increases the risk of lung cancer does not imply that a specific cancer patient who has smoked throughout her life has actually contracted the disease because of this risk factor. A coincidence is still possible.

One strategy that has been suggested in the literature is that judgments about singular causation should rely on unveiling causal mechanisms: observations of tar in the lung and genetic alterations in a cancer patient would strengthen the claim of singular causation. However, observing chains of singular events only helps with the question of singular causation if there is reason to assume that they are causally related. Thus, generic knowledge is necessary here as well (see also Danks, in press).

Since studying chain-like mechanisms does not fully solve the problem of how generic and singular causation are related, we here began with the simplest case with one cause and one effect event. In three experiments we showed that subjects used contingency information from both time series data of an individual and from group data when making judgments about singular causation.

We also showed that subjects differentiated between generic and singular causation. The responses to queries about singular causation were best explained by a variant of the SI model from Meder et al. (2014) that computes an esti-

mate of causal responsibility (Cheng & Novick, 2005). Unlike its main competitors, the SI model is sensitive to the uncertainty of both the causal structures and their parameters.

Our research is just a first step. We have compared the SI model with two power PC models. However, we kept the priors uninformative and did not model different types of priors (e.g., Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008).

Another question that we want to address in future research is how subjects assess generic causal relations in time series versus group samples. Finally, it would be interesting to broaden the scope of the SI model by applying it to more complex causal models, such as causal chains, to gain insights in the important role that mechanism knowledge plays in singular causal inferences.

Acknowledgments

We thank Jonas Nagel, York Hagmayer, and Ralf Mayrhofer for helpful comments.

References

- Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.
- Cheng, P. W., & Novick, L. R. (2005). Constraints and non-constraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*, 694–706.
- Danks, D. (in press). Singular causation. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning*. New York: Oxford University Press.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.
- 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*, 702–727.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–982.
- Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*, 277–301.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan-Kaufmann.
- Waldmann, M. R. (in press). *Oxford handbook of causal reasoning*. New York: Oxford University Press.
- Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 733–752). New York: Oxford University Press.
- Waldmann, M. R., & Mayrhofer, R. (in press). Hybrid causal representations. In B. Ross (Ed.), *The psychology of learning and motivation*. New York: Academic Press.

Appendix E

Stephan and Waldmann (2017)

Stephan, S., & Waldmann, M. R. (2017). Preemption in singular causation judgments: A computational model. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1126–1131). Austin, TX: Cognitive Science Society.

Preemption in Singular Causation Judgments: A Computational Model

Simon Stephan (simon.stephan@psych.uni-goettingen.de)

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)

Department of Psychology, University of Göttingen,
Gosserstr. 14, 37073 Göttingen, Germany

Abstract

Causal queries about singular cases are ubiquitous, yet the question of how we assess whether a particular outcome was actually caused by a specific potential cause turns out to be difficult to answer. Relying on the causal power approach, Cheng and Novick (2005) proposed a model of causal attribution intended to help answering this question. We challenge this model, both conceptually and empirically. The central problem of this model is that it treats the presence of sufficient causes as necessarily causal in singular causation, and thus neglects that causes can be *preempted* in their efficacy. Also, the model does not take into account that reasoners incorporate uncertainty about the underlying causal structure and strength of causes when making causal inferences. We propose a new measure of causal attribution and embed it into our structure induction model of singular causation (SISC). Two experiments support the model.

Keywords: singular causation; causal attribution; preemption; causal reasoning; Bayesian modeling; computational modeling

Introduction

Most people hold the belief that smoking causes lung cancer. Now, imagine that you learn that Peter, a passionate smoker, has contracted lung cancer. How strongly would you be willing to say that it was Peter's smoking that was causally responsible for his disease?

This example illustrates a scenario in which we seek an answer to a causal query about a singular case. Queries about singular causation are prevalent in everyday life and professional contexts, such as the law or medicine. How do people derive causal judgments about singular cases? Of course, the mere fact that two factors *C* and *E* are *generally* causally connected (e.g., smoking often causing lung cancer) does not necessarily imply that a *singular* or *token* co-occurrence of these events (e.g., Peter's smoking and his lung cancer) manifests a causal relationship – a singular co-occurrence might be a mere *coincidence*. On the other hand, as causality is not directly observable in the world, to what else than our general causal knowledge could we turn to obtain answers?

We are going to present a theory that builds on the idea, first formalized by Cheng and Novick (2005), that the notion of unobservable *causal powers* (Cheng, 1997) plays an essential role in singular causation judgments. Yet, we will demonstrate that Cheng and Novick's (2005) power PC model of causal attribution (CN model) makes assumptions that are not always plausible. Generally, the CN model is intended to provide a normative answer to the question how we can determine whether an observed outcome was actually caused by a potential cause factor. For example, for cases like the one above about Peter in which a potential cause *c* and an effect *e* have been observed, the CN model delivers the probability

$P(c \rightarrow e|c, e)$ with the arrow denoting a causal relation. We argue that the key problem of the model is that it treats target causes as singular causes whenever they are sufficient for the effect in a specific situation. This appears to be at first sight a reasonable assumption, yet it ignores that the exact points in time at which different causes exert their powers play an important role in singular causation judgments (see Danks, 2017): crucially, sufficient causal powers can be *preempted* by others, and in such cases they should not be held causally responsible for the occurrence of the outcome. We will argue that preemption of causes by background factors frequently occurs in singular causation scenarios, and therefore presents a problem for the CN model.

Another problem of the CN model is that it does not take into account uncertainty about both the underlying causal structure and the causal parameters (e.g., the size of the causal powers). To incorporate uncertainty about the causal parameters, Holyoak, Lee, and Lu (2010) have proposed a Bayesian version of the CN model that uses probability distributions over the parameters instead of point estimates. However, their model also neglects uncertainty about the underlying causal structure. Both sources of uncertainty have been demonstrated to influence causal learning and reasoning (Griffiths & Tenenbaum, 2005; Meder, Mayrhofer, & Waldmann, 2014). For this reason, Stephan and Waldmann (2016) proposed the structure induction model of singular causation (SISC) that incorporates both types of uncertainty. Although three experiments (Stephan & Waldmann, 2016) showed that SISC better accounted for the results than the standard power PC model of causal attribution, one shortcoming of the initial version of SISC was that it used the CN conceptualization of causal attribution that we are going to criticize in the present paper.

We will start with a theoretical section in which we defend a new measure of causal attribution as a component of SISC that is sensitive to preemption. We then present the results of two experiments. Experiment 1a confirmed that singular causation judgments deviate systematically from the predictions of the CN model in line with our revised causal attribution equation. Experiment 1b assessed participants' notion of preemption. In Experiment 2 we used a larger set of contingencies to compare the revised SISC with the CN and other models. The results of this experiment showed that both a revision of the causal attribution equation and the consideration of statistical uncertainty are crucial to explain the findings.

The Power PC Model of Causal Attribution

According to Cheng's (1997) power PC theory, causal power (or causal strength) is a hypothetical, unobservable en-

tity that represents the strength of causes. Mathematically, causal power is (in the generative case) expressed as the probability with which a target cause brings about its effect in a hypothetical world in which all alternative observed and unobserved causes of the effect are absent. Because of the possibility of unobserved alternative causes, causal power cannot be assessed directly but must be inferred based on the observed covariation and background assumptions. For generative causes, the following equation can be used to estimate the causal power w_c of a target cause C :

$$w_c = \frac{P(e|c) - P(e|\neg c)}{1 - P(e|\neg c)} = \frac{\Delta P}{1 - w_a} \quad (1)$$

In this equation, w_a represents the aggregate causal power of all alternative causes A of the effect, which are assumed to exert their influence independently of C .

Under the causal Bayes net framework, the causal power of C , w_c , corresponds to the probabilistic weight of the causal arrow that connects C with its effect E in a common effect structure in which the target cause C and the alternative causes A combine in a *noisy-OR* gate (see S_1 in Figure 2). Likewise, w_a corresponds to the weight of node A .

The CN Measure of Causal Attribution

Cheng and Novick (2005) proposed several measures of causal attribution that apply to different cases. The measure of causal attribution for cases in which both c and e are present, as in the example above about Peter, utilizes the concept of causal power in the following way to deliver the conditional probability $P(c \rightarrow e|c, e)$:

$$P(c \rightarrow e|c, e) = \frac{w_c}{P(e|c)} = \frac{w_c}{w_c + w_a - w_c \cdot w_a}. \quad (2)$$

Equation 2 shows that the CN model defines the probability with which c is causally responsible for e given that both have co-occurred by the fraction of the causal power of C and the conditional probability of the effect in the presence of C . Since the power PC theory assumes that C and A exert their causal powers independently of each other, $P(e|c)$ can be rewritten as the sum of both causal powers minus their intersection (see second step of Equation 2). Hence, what the CN model delivers is an estimation of the relative frequency of cases among all co-occurrences of C and E in which C 's causal power is sufficient for the production of the effect. Our key criticism is that this relative frequency, because it neglects the possibility of preemption, frequently overestimates the true proportion of cases in which we should actually causally attribute E 's occurrence to C .

To illustrate the problem, let us consider the results of the fictitious experiment shown in Figure 1 in which the influence of a chemical substance on the expression of a gene was investigated. As it is the case that all mice in the test group ($P[e|c] = 1$) but only one half in the control group ($P[e|\neg c] = .5$, the base rate) exhibit the gene, the results provide strong evidence for the existence of a strong effect of the chemical. In fact, by applying Equation 1 one can see that

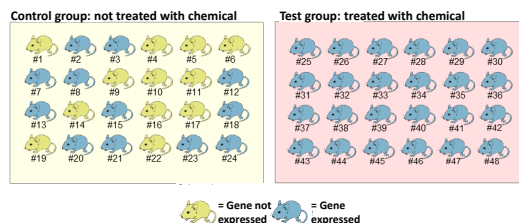


Figure 1: Illustration of a hypothetical study testing the effect of a chemical on the expression of a gene. The control group is shown on the left and the test group treated with the chemical substance on the right. Mice having the gene expressed are depicted in blue.

the causal power of the substance equals 1. Crucially, so does $P(c \rightarrow e|c, e)$. What the CN model therefore prescribes is that we should attribute causal responsibility to the chemical whenever the chemical and the gene are both present. But should we really be maximally confident that the expression of the gene in, for instance, Mouse #25 must be causally attributed to the causal power of the chemical? If you have doubts, you were probably led by a prior assumption about the point in space and time at which the causal background factors A produced the observed base rate of fifty percent: in the given scenario it seems likely that these factors (e.g., transcription factors) already produced their effects prior to the introduction of the chemical. Under this assumption, however, it seems likely that not only fifty percent of the mice in the control condition but also in the test group already possessed the gene prior to the study. Consequently, in those fifty percent of the mice it cannot be the chemical that is causally responsible for the effect because its causal efficacy has been preempted by the background factors.

A New Measure of Causal Attribution

In the example it seems appropriate to say that the expression of the gene is caused by the chemical in only about half of the observed cases in which C and E have co-occurred. This conclusion is based on the assumption that in roughly half of the cases the causal power of the chemical has been preempted by the causal power of the background factors A . We propose a new measure of causal attribution that captures this intuition by refining Equation 2 so that all cases among the joint occurrences of C and E for which the effect of C is assumed to be preempted by A are partialled out. This refined measure is given by:

$$P(c \xrightarrow{\text{singular}} e|c, e) = \frac{w_c \cdot (1 - \alpha \cdot w_a)}{w_c + w_a - w_c \cdot w_a}, \quad (3)$$

in which we introduce α as a discounting parameter that represents the assumed probability with which A is a preemptive cause of the effect. For illustration, if we assume that A has caused the observed base rate of 0.5 prior to the application of the chemical, and if we assume further that A 's causal power has produced roughly equal proportions of the effect in both groups, α takes on a value of 1. In this case, the point estimate for $P(c \xrightarrow{\text{sing.}} e|c, e)$ is about .5 instead of 1.0 that is predicted

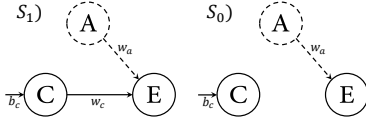


Figure 2: The two causal structures considered by the structure induction model of singular causation as mutually exclusive explanations for observed patterns of covariation. C is a general cause of E under S_1 whereas all co-occurrences of C and E are coincidental under S_0 . The parameters w_c and w_a denote the causal powers of the observed cause C and the unobserved background causes A , respectively; b_c denotes the base rate of C .

by Equation 2. The difference between the two measures is that under the CN model the presence of two sufficient causal powers (w_c and w_a) is invariably conceptualized as a case of “symmetric overdetermination”, whereas the possibility that causes can preempt each other is neglected. Equation 3 takes into account the possibility of preemption and thus delivers an estimation of the relative frequency of singular cases in which the target cause has *actually* been successful in generating the effect. In our view, preemption of C by a previously present background factor A seems to be prevalent in the cover stories typically reported in the literature (see e.g., Griffiths & Tenenbaum, 2005). However, the discounting parameter α can be set to capture other cases. For example, cases of overdetermination or cases in which A is preempted by C could be modeled by setting α to 0. In these cases, our model and the CN model make identical predictions because Equation 3 reduces to Equation 2.

SISC: The Structure Induction Model of Singular Causation

Apart from the fact that the CN model does not take into account the possibility that preempted causes should not be classified as singular causes, a further problem of the CN model is that it is insensitive to statistical uncertainty about both the underlying causal structure and the size of the causal parameters. SISC (Stephan & Waldmann, 2016) is sensitive to both types of uncertainty.

SISC was developed in the framework of causal Bayesian inference models; it takes observed data as evidence to update prior probabilities of mutually exclusive hypotheses. Under SISC, these competing hypotheses represent two causal structures that can account for a particular observed pattern of covariation. The two causal structures, S_0 and S_1 , are depicted graphically in Figure 2. While there exists a causal arrow from C to E in S_1 , which indicates that C is a general cause of E , there is no causal arrow between C and E in S_0 . Both models assume a background cause A .

The core principle of SISC can be illustrated with Figure 1. Assume someone suggests that S_0 is the causal structure that underlies the results. Under this hypothesis all observed co-occurrences of C and E would be mere coincidences. Yet, since the observed distribution of the events appears very unlikely to be coincidental, S_0 is weakened as an explanation while the alternative hypothesis, S_1 , is proportionally strengthened. In fact, the probability computed by SISC for

S_1 for the data shown in Figure 1 (i.e., the posterior probability of S_1) is almost 1. Now, imagine the same study had been conducted with a sample of merely eight mice but that $P(e|c)$ and $P(e|\neg c)$ remain the same. In this case, it seems less certain that S_1 underlies these results. Smaller samples not only increase uncertainty about the underlying causal structure, they also impede the reliable estimation of the size of parameters. SISC is sensitive to both types of uncertainty when estimating $P(c \xrightarrow{\text{sing.}} e|c, e)$.

SISC implements different steps. First, it derives the posterior probabilities for each causal structure illustrated in Figure 2. Applying Bayes’ rule, the posterior probability for a causal structure is proportional to the likelihood of the data given the causal structure, weighted by the structure’s prior probability:

$$P(S_i|D) \propto P(D|S_i) \cdot P(S_i). \quad (4)$$

$P(D|S_i)$ is the likelihood of the data given a particular structure, which is the integral over the likelihood function of the parameter values under the particular structure. $P(S_i)$ represents a structure’s prior probability. The model initially assumes that both structures are equally likely, that is, $P(S_i) = 1/2$. When data become available, the posterior for each causal structure varies systematically with the observed contingency: the higher the contingency, the more likely S_1 becomes.

Next, the model estimates the parameters b_c , w_c , and w_a , for each causal structure. To express parameter uncertainty, distributions rather than point estimates are inferred. The posterior probability distributions for the parameters, $P(w|D)$, are proportional to the likelihood of the data given the set of parameters w , weighted by the prior probability distributions of the parameters:

$$P(w|D) \propto P(D|w) \cdot P(w). \quad (5)$$

$P(D|w)$ is the likelihood of the data given the parameter values for b_c , w_c , and w_a . $P(w)$ is the prior joint probability of the parameters. The prior distributions of the parameters are independently set to flat, uninformative $beta(1,1)$ distributions. Since C does not cause E under S_0 , w_c is held fixed at 0 for this causal structure.

In the last step, SISC computes $P(c \xrightarrow{\text{sing.}} e|c, e)$ for each parameterized structure. The new discounting parameter α is set based on background assumptions about the target scenario. For the scenarios we used in the present experiments it is set to 1 because preemption seems to be highly probable. As all co-occurrences of c and e are coincidences under S_0 , $P(c \xrightarrow{\text{sing.}} e|c, e)$ is set to 0 for S_0 . For S_1 , Equation 3 is applied. The final output of SISC is a single estimate for $P(c \xrightarrow{\text{sing.}} e|c, e)$, which is obtained through integrating out the two causal structures by summing over the derived values of $P(c \xrightarrow{\text{sing.}} e|c, e)$ for each structure weighted by its posterior probability:

$$P(c \xrightarrow{\text{sing.}} e|c, e; D) = \sum_i P(c \xrightarrow{\text{sing.}} e|c, e; S_i) \cdot P(S_i|D). \quad (6)$$

Experiment 1a

The goal of Experiment 1a was to test SISC against the CN model of causal attribution for data sets with a sufficient cause, i.e., $P(e|c) = w_c = 1$, but varying base rates of the effect. Whereas the CN model predicts maximal confidence in singular causation assessments for any observed co-occurrence of C and E in this case, SISC predicts an interaction with the base rate under the assumption that A 's causal power generally preempts the effect of C . The goal of Experiment 1a was to demonstrate that this predicted deviation from the CN model is expected for the conceptual reasons discussed above. To rule out uncertainty as an explanation, we used sample sizes in our data sets for which the posterior probabilities of S_1 computed by SISC are close to 1. The predictions of the models are shown in Figure 3. We set α in Equation 3 to 1, which represents complete preemption of C by A . We also considered a Bayesian variant of the CN model that has been proposed by Holyoak et al. (2010). This model is sensitive to parameter uncertainty; it uses probability distributions over the parameters instead of point estimates. As Figure 3 shows, the predictions of both variants of the CN model converge for large sample sizes because the influence of parameter uncertainty decreases.

Methods

Participants 90 participants ($M_{age} = 33.24$, $SD_{age} = 12.50$, 35 female) were recruited via Prolific Academic (www.prolific.ac) and received a monetary compensation of £0.60.

Design, Materials, and Procedure Three contingencies (see Figure 3) were manipulated between subjects with each participant responding to two causal test queries (general causation vs. singular causation). We included the general causation query to establish that uncertainty cannot account for the predicted pattern of singular causation ratings. The task was a standard elemental causal induction task. As cover story we used the gene expression scenario (cf. Griffiths & Tenenbaum, 2005) mentioned above: subjects were asked to assume that they were biologists who are interested in whether a particular chemical causes the expression of a particular gene in mice. Subjects read that they will be asked to conduct an experiment on the computer screen in which they will treat a random sample of mice with the substance while a control sample will remain untreated. It was mentioned that the control sample is important as some individuals may show the gene expression for other reasons.

Participants were presented with an interactive animation showing the two samples arranged as in Figure 1, and a pipette containing a reddish chemical substance. All mice had gray color in this animation. Participants then dropped the substance into the test group area, whereupon the background color changed to a light red. On the next screen, subjects checked the results of the experimental manipulation by dragging a small magnifying glass over all the mice. Mice with the gene then became blue and those without became yellow. The final state of the animation looked like Figure 1.

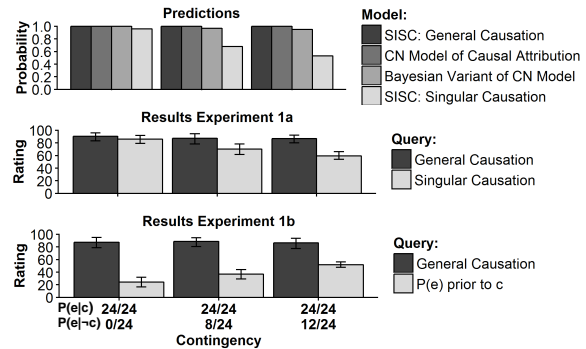


Figure 3: Model predictions and results of Experiments 1a and b. The results show mean ratings and 95% bootstrapped CI s. Dark bars show general causation judgments; light bars singular judgments.

Subsequently, participants responded to two test questions. The general causation query referred to the causal *structure*. Participants were asked to indicate on a slider how confident they were that the chemical has an effect on the expression of the gene (from “very certain that the chemical has no effect” to “very certain that the chemical has an effect”). The singular causation query asked subjects about Mouse #25 from the test group. Participants were asked to indicate on a slider how confident they were that it was the chemical substance that caused the expression of the gene in this single case (from “very certain that it was not the chemical” to “very certain that it was the chemical”).

Results and Discussion

Figure 3 shows the results. The prediction for general causation responses corresponds to the posterior probability of S_1 computed by SISC. As predicted by the posterior probability of S_1 , all general causation ratings were high, indicating very little uncertainty about the general causal structure. The singular causation ratings, by contrast, decreased with an increasing base rate of the effect, as predicted by SISC but not by the two CN models. The results of a multilevel model analysis revealed significant main effects for type of causal query, $\chi^2(1) = 32.45$, $p < .001$, as well as contingency, $\chi^2(1) = 12.63$, $p < .01$. General causation ratings were, on average, higher than singular causation ratings. Figure 3 shows that the main effect of contingency is driven by the decrease in singular causation ratings. Planned contrasts revealed that the general causation ratings neither differed between the first and second contingency, $t(80) = 0.60$, nor between the second and third contingency, $t(80) = 0.13$. Consequently, the interaction effect of query \times contingency was also significant $\chi^2(1) = 13.10$, $p < .01$. Planned contrasts breaking down this interaction effect showed that the difference between general and singular causation ratings was higher for the second than for the first contingency, $t(80) = 2.10$, $p < .05$, $r = .23$, and also higher for the third compared to the second contingency, $t(80) = 3.70$, $p < .001$, $r = .38$. In sum, both the trends for general as well as for singular causation ratings are captured well by SISC.

The observed trend for the singular causation judgments is, however, neither predicted by the CN model using point estimates nor by the Bayesian extension incorporating parameter uncertainty.

Experiment 1b

The goal of Experiment 1b was to assess how likely participants think it is that a particular individual from the test group already exhibited the effect caused by the background factors prior to the occurrence of the cause. Thus, instead of singular causation judgments for a particular individual, we asked subjects to provide a probability judgment. Crucially, responses to this query provide us with an estimate of the α value in Equation 3 that participants assumed.

Methods

Participants 88 participants ($M_{age} = 31.22$, $SD_{age} = 10.84$, 42 female) participated in this only study and received a monetary compensation of £ 0.60.

Design, Materials, and Procedure The study design and the materials were the same as in Experiment 1a. The only difference was that, instead of a singular causation judgment for Mouse #25, we asked participants how likely they think it is that this individual already had the gene expressed prior to the experiment. The general causation query remained the same.

Results and Discussion

Figure 3 shows that we replicated the pattern for general causation judgments found in Experiment 1a. Planned contrasts revealed that these ratings did not differ (all t values < 1). However, the probability judgments about the presence of the effect prior to the application of the chemical in the single case showed the opposite trend as the singular causation judgments in Experiment 1a. This finding supports our hypothesis that assumptions about preemption influence singular causation judgments, as predicted by Equation 3. Planned contrasts confirmed that ratings increased from the first to the second, $t(71) = 2.67$, $p < .01$, $r = .30$, and also from the second to the third contingency, $t(71) = 3.16$, $p < .01$, $r = .35$. Furthermore, the results indicate that participants indeed assumed high α values.

Experiment 2

Experiment 1a showed that singular causation ratings for sufficient causes deviate systematically from the predictions of the CN models. This deviation is predicted as a consequence of assumptions about preemption relations between C and A . Experiment 2 pursued two main goals: first, we aimed to test SISC using a larger set of contingencies with a combination of different levels of $P(e|c)$ and $P(e|\neg c)$. Second, we wanted to demonstrate that parameter and structure uncertainty indeed influence general and singular cause judgments. We used the set of contingencies studied in Buehner, Cheng, and Clifford (2003) but excluded the one contingency from the set in which the effect never occurs. It does not make sense to ask for singular causation if the effect is absent. The data sets and model predictions are shown in Figure 4. We set

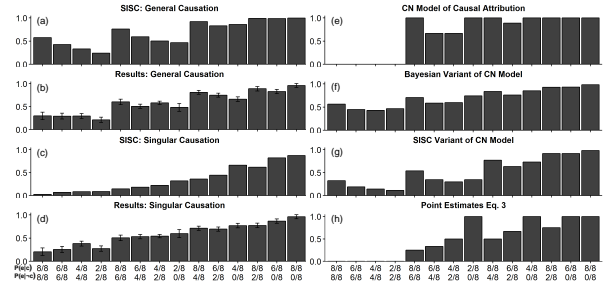


Figure 4: Predictions of different models and results (means and within-subjects adjusted 95% CIs) of Experiment 2. Graphs (a) and (b) refer to general causation assessments. All other graphs refer to singular causation assessments.

the discount parameter α to 1 again.

Methods

Participants 82 participants ($M_{age} = 34.41$, $SD_{age} = 10.42$, 31 female) participated in this online study and were paid £ 1.00 for their participation.

Design, Materials, and Procedure The causal query (general causation vs. singular causation) was manipulated between subjects, whereas contingency was varied within subject. The fourteen contingency data sets were presented in random order. We used the same cover story as in Experiment 1, except that subjects read that they will investigate the effects of fourteen different chemicals on fourteen different genes in fourteen different samples. We pointed out that the results of the studies are independent of each other. The assignment of mice to the cells of the contingency tables was randomly determined. Also the test mouse for the singular query showing both c and e was randomly chosen prior to the experiment.

Results and Discussion

Figure 4 shows the results and the predictions of the different models: (a) and (b) display the predictions of SISC for general causation and the mean general causation responses. Panels (c) and (d) show the predictions of SISC and the results regarding the singular causation queries. Predictions of the standard CN model and its Bayesian variant are displayed in (e) and (f). Graph (g) shows predictions of SISC when α is set to zero but both structure and parameter uncertainty are incorporated. Finally, (h) shows point estimates of Equation 3 while neglecting statistical uncertainty.

Table 1: Model comparisons for singular causation judgments in Experiment 2. ΔP refers to the different contingency levels (.00, .25, .75, 1.00) within the whole data set; $r_{\Delta P}$ expresses the model fits for these levels. N/A represents undefined values.

Fit measure	SISC	CN Model	Bayesian CN Model	SISC CN Model	Point Est. Eq. 3
$r_{\Delta P=.00}$.72	N/A	-.78	-.68	N/A
$r_{\Delta P=.25}$	1.00	.21	.38	-.61	.98
$r_{\Delta P=.50}$.88	.68	.79	.44	.85
$r_{\Delta P=.75}$	1.00	N/A	1.00	-1.00	1.00
$M_{\Delta P}$.90	.44	.35	-.46	.94
$r_{overall}$.94	.88	.93	.90	.90
R^2	.88	.77	.87	.82	.82
$RMSE$.11	.27	.09	.14	.22

The overall pattern for both general and singular causation ratings was captured best by the revised version of SISC. As in our previous research, the results show that participants differentiated between general and singular queries. Moreover, the responses to the general causation query replicate those found in Griffiths and Tenenbaum (2005). Most importantly, the singular causation assessments were captured best by the revised SISC. The other models, by contrast, struggled to account for the local trends observed within the subsections of the contingency set in which ΔP is constant. The difference between the singular causation ratings and the point estimates for the revised causal attribution measure (Equation 3) implies that participants were sensitive to structure and parameter uncertainty.

A multilevel model analysis confirmed the main effect of contingency, $\chi^2(13) = 1010.04$, $p < .001$, as well as the interaction between contingency \times query, $\chi^2(13) = 49.39$, $p < .001$, that is shown in Figure 4. To test the different models, we computed different fit measures shown in Table 1. As can be seen there, SISC achieved a good fit in the overall fit measures (bottom part of the table). It explained most variance, with $R^2 = .88$, and yielded the second smallest *RMSE* of .11. Yet, all models obtained relatively high values on the global measures. Even the CN model with the lowest overall fit accounted for 77 percent of the variance. The similarity between the models is not unexpected, however, as all models are sensitive to ΔP . More interesting are the fit measures for the subsections of the contingency set in which ΔP is kept constant. The upper part of Table 1 shows that SISC yielded high fit values there, too, and hence accounted well for these local trends, whereas the Bayesian CN model, which yielded the smallest *RSME*, even showed negative correlations here.

General Discussion

We addressed two different problems that the power PC framework of causal attribution (Cheng & Novick, 2005) faces: first, the CN model attributes causal responsibility for the occurrence of a particular effect e to a present singular event c whenever its causal power is sufficient to bring about the effect. We have argued that this conceptualization fails to take into account that people make assumptions about the point in time at which different causal powers exert their influences. Not every manifestation of a sufficient cause c needs to be causally responsible for an observed outcome; it might be the case that a competing cause (e.g., a) preempts it. This problem of *redundant* causation, which occurs whenever two causes are individually sufficient for the effect, is widely acknowledged in the philosophical literature as a challenge for models of causation (see, e.g., Paul & Hall, 2013). To account for the possibility of preemption we have modified the equation developed by Cheng and Novick (2005) as an account of causal attribution. The revised equation includes the discount parameter α that can be set to express domain-related assumptions about the temporal relations between the alternative causal factors. A second shortcoming of the standard causal attribution model (Cheng & Novick, 2005) is that it

does not take into account statistical uncertainty about structure and causal parameters (cf. Griffiths & Tenenbaum, 2005; Meder et al., 2014). Our model SISC remedies both shortcomings. It is sensitive to both the temporal relations between the alternative causes and to statistical uncertainty. Our experiments showed that both aspects are important to account for subjects' judgments about singular causation.

We have set the discount parameter α to 1 in Equation 3 which implies a complete preemption relation between A and C whenever A 's causal power is sufficient in a situation. Better fits might be possible by estimating the size of α for each individual subject separately. We avoided this strategy to demonstrate that model improvements can already be achieved with very general assumptions. The goal of future experiments will be to manipulate the size of α by manipulating domain assumptions about the temporal relations between C and A . Cases in which α is 1 are situations in which A always preempts C . The cover stories used in the present experiments are an example in which it is plausible to assume that A represents a temporally stable factor that has already been efficacious prior to the manipulation of C . Although preemption seems to be the default situation in most singular causation scenarios, there might be rare cases in which other assumptions need to be made. Consider cases of symmetric overdetermination that have also been discussed in the literature (see Paul & Hall, 2013): in the famous firing squad scenario, for example, in which each shooter is a sufficient cause for the death of the target, a possible intuition is that each shooter should be counted as a singular cause of the death of the victim. In this case, alpha would have to be set to zero. Similarly, alpha would have to be set to zero if C preempts A so that A cannot manifest its potential causal power. Cases of temporal variability between C and A might also be an interesting topic for future studies.

Acknowledgments We thank Jonas Nagel and Ralf Mayrhofer for helpful discussions.

References

- Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1119.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405.
- Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, 112, 694–706.
- Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 201–215). New York: Oxford University Press.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 334–384.
- 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, 702–727.
- Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, 121, 277–301.
- Paul, L. A., & Hall, E. J. (2013). *Causation: A user's guide*. Oxford University Press.
- Stephan, S., & Waldmann, M. R. (2016). Answering causal queries about singular cases. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2795–2801). Austin, TX: Cognitive Science Society.

Appendix F

Stephan, Mayrhofer, and Waldmann (2018)

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2018). Assessing singular causation: The role of causal latencies. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1080 – 1085). Austin, TX: Cognitive Science Society.

Assessing Singular Causation: The Role of Causal Latencies

Simon Stephan (sstepha1@gwdg.de) Ralf Mayrhofer (rmayrho@gwdg.de)

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)

Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Singular causation queries require an assessment of whether a singular co-occurrence of two events c and e was causal or simply coincidental. The current study builds on our previous research (Stephan & Waldmann, 2018) in which we proposed a computational model of singular causation judgments. The model highlights that singular causation judgments need to take into account the power of the target cause C and of alternative causes A , as well as the possibility of preemption. What was missing was a detailed model allowing us to estimate the probability of preemption of a target cause by the alternative causes. The present research fills this gap by elaborating the temporal assumptions that might enter assessments of singular causation. We focus on assumptions about temporal precedence between target and alternative causes, with a specific focus on assumptions about causal latency. We report the results of two new experiments supporting the model.

Keywords: singular causation; causal attribution; preemption; time; causal reasoning; computational modeling

Causal representations support various inferences. They enable us to make predictions, to form diagnoses, or to make judgments about singular causation (Waldmann, 2017). In the present research our goal is to develop and test a model explaining judgments about singular causation. How does a person come to believe, for example, that it was the medicine she took that caused her to feel sick, or that it was the combination of keys she pressed that caused the laptop's screen to turn dark, or that it was the storm last night that caused the flower pot to be shattered into pieces? To put it more generally, how do reasoners assess whether a singular instantiation of a cause factor C was actually causally connected to a singular instantiation of its effect E ?

Judgments about singular causation are so prevalent in everyday life that it is easily missed that validating them is a challenging task for the mind. The reason why it is cognitively challenging is that causal powers that bind some events together are not directly accessible to our senses (Cartwright, 1989; Cheng, 1997). The computational problem that needs to be solved is how a genuine *causal co-occurrence* of events can be discriminated from a mere *coincidental* one.

Stephan and Waldmann (2018) have proposed a computational model intended to provide a solution to this problem. Their model is a generalization of Cheng and Novick's (2005) power PC model of causal attribution, which itself relies on Cheng's (1997) power PC theory. Cheng (1997) has shown how the unobservable powers of causes, operationalized as the probability with which causes generate their effects in the hypothetical absence of alternative causes, can be inferred from observable covariation data. Cheng and Novick (2005) adopted this framework and have applied it to the question of how causal power knowledge ought to be used to make

causal attributions in different contexts. For example, when it is known that a potential cause C and a potential effect E have co-occurred on an occasion (c, e), their model provides an answer to $P(c \rightarrow e|c, e)$, the probability that c and e were causally connected on this occasion.

Stephan and Waldmann (2018) criticized and refined Cheng and Novick's (2005) model. Cheng and Novick's model focuses solely on the causal powers of the target and the alternative causes, embodying the assumption that a singular instantiation of C and E is less likely to be coincidental when C is known to operate with a large causal power. However, the model neglects another possibility why a singular co-occurrence of C and E might have been coincidental: causes, no matter how strong their power is, can be *preempted* in their efficacy by competing alternative causes (for an overview of theories on preemption, see, e.g., Paul & Hall, 2013). It is possible that an alternative cause intercepts the target cause and generates the effect before the target cause has had a chance. Stephan and Waldmann (2018) have therefore proposed a refined model that includes a term that captures the probability of preemption through alternative causes.

To estimate the probability that a target cause was preempted, what needs to be considered beyond the powers of the potential causes is temporal information. One type of temporal information that should influence how strongly a reasoner believes that a target cause was preempted by an alternative cause is the assumed difference between their instantiation times: everything else being equal, a target cause is more likely to be preempted by an alternative cause if the latter occurs earlier than the former. Stephan and Waldmann (2018) reported a set of experiments in which the cover stories suggested that unobserved alternative causes occurred prior to the target cause, and participants' singular causation judgments were explained well by the modified model incorporating the possibility of preemption.

Another type of temporal information that is likewise relevant to assess the probability of preemption is information about *causal latency*, by which we mean the time it takes a cause to produce its effect. Consider a situation in which a potential alternative cause A is instantiated simultaneously with or even later than the target cause C . In such situations, c can still be preempted by a when a 's latency is shorter than c 's.

Stephan and Waldmann (2018) did not spell out in their article how causal latency information about the competing potential causes of an outcome can be formally represented and combined with causal power information to es-

timate $P(c \rightarrow e|c, e)$ and the probability of preemption. Nor did they manipulate causal latency information in their experiments. We begin to address these shortcomings in the present research, and will primarily focus on the role causal latency plays for singular causation judgments. First we will review Cheng and Novick’s (2005) model and contrast it with Stephan and Waldmann’s (2018) modified version.

The Power PC Model of Causal Attribution

Cheng and Novick (2005) proposed a model in which the probability with which an instantiation c of a cause factor C has actually caused a token event e instantiating an effect factor E is represented by $P(c \rightarrow e|c, e)$. To estimate this probability, they have made use of Cheng’s (1997) causal power theory.

Cheng (1997) has shown how the power of a cause factor C , denoted by w_C , can be estimated from observable covariation data given a number of causal background assumptions. A graphical representation of the unobservable causal structure that is assumed by causal power theory to underlie the observable contingency ΔP between a target cause C and a target effect E is shown in Fig. 1 (see also Griffiths & Tenenbaum, 2005; Pearl, 2000). The theory considers two causal influences on E : the target cause C , and A , with A representing the sum of all unobserved alternative causes of E . It is assumed that C and A occur independently of each other with the base rates b_C and b_A . Furthermore, C and A are assumed to cause E with independent (i.e., non-interacting) powers, denoted by w_C and w_A , respectively. These assumptions allow it to explain the probability of E : $P(E) = b_C \cdot w_C + b_A \cdot w_A - b_C \cdot w_C \cdot b_A \cdot w_A$. Accordingly, the probability of E given C is $P(E|C) = w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A$, and the base rate of E is $P(E|\neg C) = b_A \cdot w_A$. The latter two equations provide a causal explanation of the observed contingency: $\Delta P = w_C + b_A \cdot w_A - w_C \cdot b_A \cdot w_A - b_A \cdot w_A$. Substituting $b_A \cdot w_A$ with $P(E|\neg C)$ and re-arranging the equation, one obtains the causal power of C :

$$w_C = \frac{\Delta P}{1 - P(E|\neg C)}. \quad (1)$$

As an illustration, consider the table in Fig. 1 with the following entries: $n(c, e) = 21$, $n(c, \neg e) = 3$, $n(\neg c, e) = 12$, and $n(\neg c, \neg e) = 12$. Imagine these data resulted from a study testing whether a drug causes nausea as a side effect. In the control group ($\neg c$), 12 of 24 subjects developed nausea. The causal power theory assumes that these cases are due to the unobserved factors included in A . In the treatment group 21 subjects had nausea. As A is supposed to occur equally likely in C ’s presence, these cases are explained by the joint influence of C and A . Based on the independence assumption we can infer that 12 of the 21 subjects with nausea would have developed nausea due to A alone had C not been present. Thus, there remained 9 subjects in which C had the chance to reveal its power. As there are 21 subjects with nausea, we can therefore conclude that the drug *exclusively* caused nausea in nine of these 12 cases. Its causal power thus is $w_C = .75$.

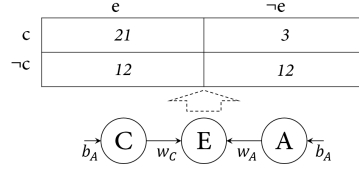


Figure 1: The relation between observable covariation data and unobservable causal structure. C and E in the causal structure denote the cause and effect factor, respectively. A comprises all unobserved alternative causes of E . b_C and b_A denote the base rates of C and A ; w_C and w_A denote the causal powers of C and A .

Now imagine you learned that a person took the drug (c) and felt sick (e). To compute $P(c \rightarrow e|c, e)$, Cheng and Novick (2005) proposed the following equation:

$$P(c \rightarrow e|c, e) = \frac{w_C}{w_C + w_A - w_C \cdot b_A \cdot w_A} = \frac{w_C}{P(E|C)}. \quad (2)$$

In this equation the power of C is in the numerator and the conditional probability of E given C is in the denominator. Stephan and Waldmann (2018) argued that what this equation thus estimates is the relative frequency of cases among all observed co-occurrences of C and E in which C ’s power was probabilistically sufficient for E . To see this, consider first how often E occurred in the presence of C in our example. This was the case in 21 of the 24 test group subjects. Hence, we obtain $P(E|C) = .875$. From Eq. 1 we know that C has a power of $w_C = 0.75$, implying that C was sufficient in 75 percent of the treatment subjects (in 18 of the 24). Applying Eq. 2, we see that model thus concludes that C caused E in 18 of the 21 cases in which C and E co-occurred, yielding $P(c \rightarrow e|c, e) = \frac{18}{21} = 0.86$. The probability of a singular co-occurrence having been causal is supposed to be 86 percent.

A Modified Model Sensitive to Preemption

Stephan and Waldmann (2018) have argued that the power PC model of causal attribution tends to overestimate the probability of singular causation. Their argument was that not on every occasion on which a cause factor C is probabilistically sufficient to generate E it has actually caused E because it is possible that C has been preempted by an alternative cause factor A that succeeded in causing E before C could have taken effect. To illustrate the problem, imagine the extreme case in which all alternative factors A generated their effects before C . In this case we would say that there are 12 subjects with nausea due to A in each group, and it seems natural to say in this situation that C actually caused nausea only in those nine subjects that were added to these 12. The probability that a subject from the test group suffered from nausea due to the drug would hence be $P(c \rightarrow e|c, e) = \frac{9}{21} = 0.43$, and not 0.86. The power PC model of causal attribution seems to yield the correct result only in a situation in which C produces all its effects prior to A , or when the two causes are in a relation of *symmetric* overdetermination.

To incorporate the possibility of preemption by alternative causes, Stephan and Waldmann (2018) have proposed the following refined equation:

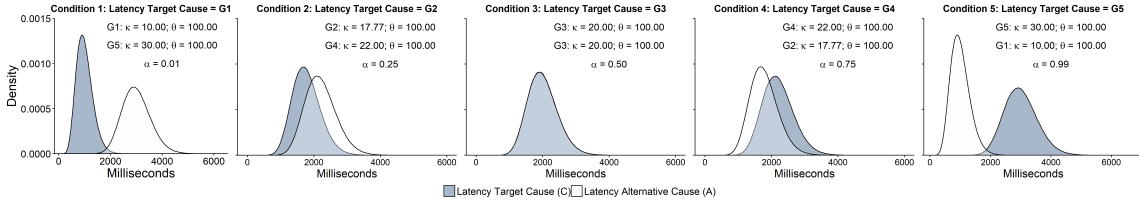


Figure 2: Pairs of gamma distributions contrasted in the five conditions of Exp. 1. The shape (κ) and scale (θ) parameters of the five different gamma distributions are listed for each pair. The depicted α values were obtained using the MC algorithm corresponding to Eq. 4. The dark distributions show the causal latencies of the target cause and the light distributions show the causal latencies of the alternative cause.

$$P(c \rightarrow e|c, e) = \frac{w_C - w_C \cdot b_A \cdot w_A \cdot \alpha}{P(E|C)} = \frac{w_C \cdot (1 - b_A \cdot w_A \cdot \alpha)}{P(E|C)}. \quad (3)$$

Eq. 3 extends the numerator of Eq. 2 by subtracting from w_C the product of w_C , b_A , w_A , and a newly introduced parameter α . The product of b_A , w_A , and α captures the probability of preemption. The intersection of w_C , b_A , and w_A identifies the occasions in which C and A are both probabilistically sufficient to generate E . This part of the term is relevant because the problem of preemption occurs only on occasions in which the potential causes C and A are both probabilistically sufficient to generate the effect. On those occasions it can either be the case that C preempts A , that A preempts C , or that both act synchronously (discussed in the philosophical literature under the term *symmetric overdetermination*, see, e.g., Paul and Hall 2013). Stephan and Waldmann (2018) introduced the α parameter as a weighting factor that narrows down the intersection of w_C , b_A , and w_A to those occasions on which C is preempted by A . To illustrate the idea, consider again the nausea example. In this example, $w_C \cdot b_A \cdot w_A = 0.75 \cdot 0.50 = 0.375$. Now imagine again the extreme case in which the factors in A produced their effects prior to C . In this scenario, all 37.5 percent of the cases in which C and A were both sufficient for E were actually caused by A . Such a situation can be modeled by setting α to 1. This would yield $P(c \rightarrow e|c, e) = \frac{0.75 - 0.375 \cdot 1.0}{0.875} = 0.43$. By contrast, in the extreme cases in which C generates its effects prior to A or synchronously with A , α should be set to 0. In this case Eq. 3 reduces to Eq. 2.

Incorporating Causal Latency Information

We showed in the previous sections how the possibility of preemption can be incorporated into a model of causal attribution. The open question is how α can be estimated in different contexts. This is where temporal information about the potential causes comes into play.

We will focus on the situation in which A only consists of a single alternative cause factor of the effect. As pointed out above, an obvious factor influencing α in such situations is the difference between the onset times of C and A . This difference can be represented by $\Delta_t = t_a - t_c$. Everything else being equal, when A occurs earlier than C , it is more likely that A preempts C than vice versa.

Additionally, the probability of preemption is influenced by the causal latency of the potential causes. By causal latency we mean the time it takes a cause to produce its effects. Variation in the latency with which a cause generates its effect opens up the possibility that even when C and A are instantiated simultaneously, or when A is instantiated later than C , C could still be preempted by A .

To model causal latencies we will use gamma distributions, which are, for example, used in queueing theory to model waiting times. In recent studies, Bramley, Gerstenberg, Mayrhofer, and Lagnado (in press) used gamma distributions to model the role of time in causal structure induction (see also Lagnado & Speekenbrink, 2010). A gamma distribution is a continuous probability distribution characterized by two parameters: shape, $\kappa > 0$, and scale, $\theta > 0$. The expected value of a random variable X following a gamma distribution is $E[X] = \kappa \cdot \theta$. Its variance is $Var[X] = \kappa \cdot \theta^2$. Different pairs of gamma distributions that we contrasted in our experiments are depicted in Fig. 2.

The representation of causal latencies based on gamma distributions can be used to estimate α in different types of situations. We will here focus only on situations in which it is known that C , A , and E are all present. Moreover, the causal latency distributions of C and A and the size of Δ_t are known. In these situations α corresponds to:

$$\alpha = P(t_{A \rightarrow E} + \Delta_t < t_{C \rightarrow E} | e, c, a, \Delta_t). \quad (4)$$

In this equation, $t_{C \rightarrow E}$ and $t_{A \rightarrow E}$ denote the causal latencies of C and A (cf. Bramley et al., in press), which are given by the respective gamma distributions. In this situation α corresponds to the probability that the sum of the causal latency of the competing cause A and the time lag between C 's and A 's onset times is smaller than the causal latency of C , given e , c , a , and Δ_t . This probability can be estimated with the following Monte Carlo (MC) algorithm:

1. Sample N pairs of causal latencies (x_c, x_a) from the gamma distributions of C and A , respectively.
2. Calculate $x_{a'} = x_a + t_a - t_c$ for all sampled x_a -values.
3. Count all pairs for which $x_c > x_{a'}$.
4. Divide this count by N .

The values for α depicted in Fig. 2 were obtained by applying this algorithm with $N = 10,000$.

Experiment 1

The goal of Exp. 1 was to compare different scenarios intended to manipulate α and to compare the predictions of our model for these situations with people’s singular causation judgments. To simplify the task, and to isolate the influence of causal latency, we used a test scenario in which the competing cause factors occurred simultaneously (i.e., $\Delta_t = 0$). Furthermore, we considered only deterministic causes in this first experiment. Fig. 2 shows the five gamma distributions (G1 - G5; higher numbers indicate higher expected values) that we contrasted in five conditions. The latency distributions belonging to the target cause C in each condition are depicted in dark blue. For the first pair, for example, in which we contrasted G1 and G5, the target cause factor’s latency follows G1, whereas the latency of the alternative cause follows G5. Fig. 2 also shows the different α values estimated with the algorithm presented above. For the first pair $\alpha = 0.01$, which in the case of deterministic causes directly corresponds to the probability that C was preempted by A . Thus, participants should be confident in this condition that it was indeed the target cause C that brought about the observed outcome. In the fifth condition, by contrast, in which C followed G5 and A followed G1, participants should be confident that C did not cause the effect. Fig. 2 also shows that all the other conditions should elicit more uncertainty. In the third condition, for example, in which C and A have the same latency distribution (G3), $\alpha = 0.50$. Here participants should be maximally uncertain about the singular cause of the outcome.

Fig. 4 shows the predictions of the power PC model of causal attribution (Eq. 2) as well as those of our refined model that computes α according to the algorithm corresponding to Eq. 4. We labeled this model “Alpha Precise” because we also considered an “Alpha Coarse” model. Alpha Coarse estimates only whether the expected values of the competing distributions differ and neglects their variances. It thus assigns a value of 1 to α if the expected value of the causal latency distribution of the target cause is larger than the one of the competing cause, and 0 otherwise. Alpha Coarse treats the situation in which both causes follow identical latency distributions as a case of *symmetric overdetermination*, and therefore predicts that both cause factors should be seen as singular causes in this situation.

Participants

Two hundred subjects ($M_{age} = 27.14$, $SD_{age} = 8.71$, 119 females) who had at least an A-level degree and who were native English speakers were recruited via Prolific (www.prolific.ac). Subjects were paid £0.80 for their participation.

Design, Materials, and Procedure

Participants were randomly assigned ($n = 40$) to one of five key conditions. These conditions varied with respect to the contrasted gamma distributions and with respect to the gamma distribution associated with the target cause (see Fig. 2). Because we included several balancing factors, the

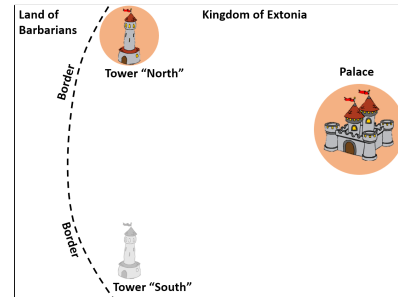


Figure 3: Illustration of the learning task used in the experiments.

full design was a $2 \times 2 \times 2 \times 5$ between-subjects design. The additional balancing factors will be introduced below.

Subjects were presented with a scenario about a fictitious medieval kingdom called “Extonia”. They read that the king had two watchtowers (“North” and “South”) built at the border to protect his empire from barbarians. These towers were instructed to send carrier pigeons to the palace to cause alarm whenever barbarians are spotted. Participants were then asked to take the perspective of Extonia’s secretary of defense who routinely inspects the flight durations of the pigeons from the two towers. Participants read that the flight durations tend to differ between different pigeons, and that they will therefore observe a sample of thirteen pigeons from each tower. Before participants could proceed to the learning task, they had to pass an instruction check.

During the learning task the screen looked similar to the picture in Fig. 3. Whether participants began with tower “North” or “South” was balanced between subjects. In each trial the sending of a carrier pigeon was indicated with a delay of 500ms by a circling of the watchtower. The arrival of the pigeon was indicated by a colored circle surrounding the palace. The thirteen flight durations for each tower corresponded to thirteen quantiles of the respective gamma distribution. We used quantiles so that subjects could be presented small but yet representative samples. After each trial, participants had to click a “Next” button, which was operational 500ms after the circle around the palace had been displayed. The flight durations were presented in random order.

The test scenario described a singular situation in which the palace had been alarmed by a tower so that it was possible to repel a horde of barbarians. Participants read that the people of Extonia wanted to decorate the tower that was responsible for the alarm but that there was the problem that both towers had actually sent their pigeons simultaneously. To express their opinion about which tower was the actual cause of the alarm, participants indicated on an eleven-point rating scale (end points: “Definitely not caused by Tower ‘North/ South’” and “Definitely caused by Tower ‘North/ South’”; midpoint “50:50”) how strongly they believed that the alarm was caused by Tower “North”/ “South”. Whether the target cause was tower “North” or “South” was balanced between subjects. The orientation of the rating scale was also balanced between subjects.

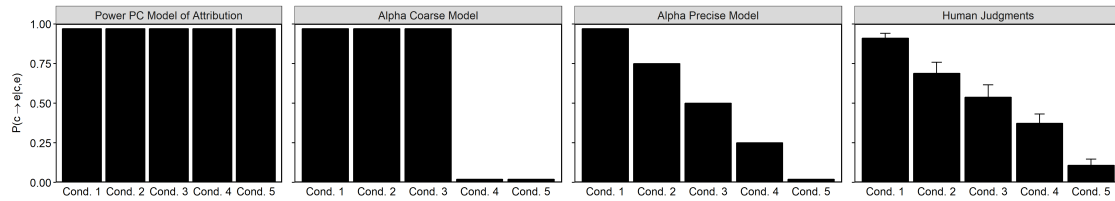


Figure 4: Model predictions and results (mean singular causation ratings and 95% bootstrapped CIs) for the different conditions of Exp. 1. The predictions of the Power PC Model were obtained from Eq. 2. The predictions of the Alpha Coarse Model are based on α being obtained through an ordinal ranking of the expected values of the compared gamma distributions. For the predictions of the Alpha Precise Model α was calculated with the MC method corresponding to Eq. 4.

Results and Discussion

The results are shown in Fig. 4. Participants’ singular causation ratings ($M_1 = 0.91$, $SD_1 = 0.10$, $M_2 = 0.69$, $SD_2 = 0.22$, $M_3 = 0.54$, $SD_3 = 0.24$, $M_4 = 0.37$, $SD_4 = 0.19$, $M_5 = 0.11$, $SD_5 = 0.12$, from left to right) followed a negative linear trend. A polynomial trend analysis confirmed the negative linear trend, $F(4, 195) = 111.70$, $p < .001$, $r = .83$. No other polynomial trend was significant.

These results are at odds with the predictions made by the power PC model of causal attribution (Eq. 2). Subjects took into account information about the causal latency of the potential cause factors to derive singular causation judgments. Moreover, the singular causation judgments followed the predictions of the Alpha Precise model, which utilizes gamma distributions to represent causal latencies. The correlation between model predictions and results was high, $r = .99$, and statistically significant $t(3) = 14.88$, $p < .001$.

Experiment 2

In Exp. 1 we tested deterministic causes because we aimed to isolate the influence that causal latency exerts on singular causation judgments. As Eq. 3 shows, however, the probability of preemption is given by the product of the causal powers and α . In Exp. 2 we therefore studied probabilistic causes. We tested a scenario in which both causes either had a causal power of $w_C = w_A = 0.83$ or of $w_C = w_A = 0.5$. Additionally, we manipulated α by using the first pair of gamma distributions (G1 vs. G5) shown in Fig. 2. We tested this combination of causal power and latency because it leads to an interesting interaction effect, depicted in Fig. 5. When the target cause’s latency is high (G5), our model predicts that ratings in the high-power condition should be lower than in the low-power condition. One reason is that the product that is subtracted from w_C in the high-power condition is so large that the numerator becomes smaller ($0.83 - 0.83 \cdot 0.83 \cdot 0.99 = 0.15$) than in the low-power condition ($0.5 - 0.5 \cdot 0.5 \cdot 0.99 = 0.25$). Furthermore, the denominator is smaller in the low-power condition than in the high-power condition ($0.5 + 0.5 - 0.5 \cdot 0.5 = 0.75$ vs. $0.83 + 0.83 - 0.83 \cdot 0.83 = 0.97$), which means that the numerator in the low-power condition is increased more strongly than in the high-power condition. When target cause’s latency is low (G1), by contrast, the product that is subtracted in the numerators becomes small, and our model predicts that we should see a reversed order of judgments. Fi-

nally, our model also predicts a main effect of causal latency: causes that tend to precede the efficacy of their competitors should receive higher singular causation ratings than causes whose competitors tend to preempt them.

Fig. 5 also shows the predictions of the power PC model (Eq. 2). As this model is blind to causal latency information, it predicts only a main effect of causal power. Our model does not predict a main effect of causal power.

Participants

One hundred and sixty subjects ($M_{age} = 38.14$, $SD_{age} = 12.20$, 116 females) who had at least an A-level degree and who were native English speakers were recruited via Prolific (www.prolific.ac). They were paid £ 1.20 for participation.

Design, Materials, and Procedure

Subjects were randomly assigned to one of four conditions ($n = 40$) that resulted from a 2 (causal power: $w_C = w_A = 0.83$ vs. $w_C = w_A = 0.50$) \times 2 (causal latency: G1 vs. G5) between-subjects design.

The materials and procedure were largely identical to those of Exp. 1, with the following exceptions: first, we added information about the probabilistic causal nature of the towers to the instructions. Participants read that pigeons might get lost on their way to the palace and that it would hence be important to learn the pigeons’ arrival rates. Secondly, the learning task was modified such that causal power information could be conveyed. Other than in Exp. 1, we showed subjects 24 pigeons per tower, to ensure that all participants observe a sufficient number of “successful” pigeons to be able to learn the causal latencies. Subjects in the high-power condition observed 20 successful pigeons per tower, whereas subjects in the low-power condition observed 12 successful pigeons per tower. The flight durations corresponded to 12 or 20 percentiles of the respective latency distributions. Whenever a pigeon failed to reach the palace, the words “Pigeon probably lost” were displayed five seconds after the pigeon had been sent out, which corresponded to the 99.9th percentile of G5.

The test questions were identical with the ones in Exp. 1. Additionally, on a separate screen we asked participants to estimate the causal powers of the towers, as we wanted to control for the possibility that subjects’ representations of causal power might be influenced by causal latency. For example, the tower “North” participants were asked the following question: “Based on what you have learned: how many out of 10

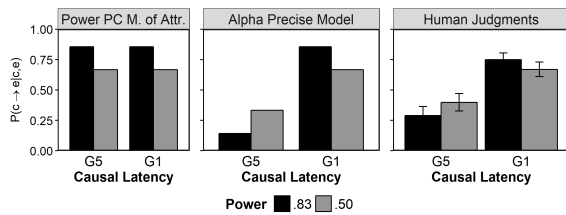


Figure 5: Model predictions and results (mean singular causation ratings) of Exp. 2. Error bars denote 95% bootstrapped CIs.

letter pigeons sent from Tower ‘North’ would make it to the palace?”. Ratings were provided on an eleven-point scale (0 to 10). The test questions were presented in counterbalanced order.

Results and Discussion

The results are summarized in the right panel of Fig. 5. The singular causation ratings followed the pattern predicted by our model, whereas the power PC model of causal attribution cannot explain the results. When α was high, ratings in the low-causal power condition were higher than those in the high-causal power condition. The reversed pattern was obtained when α was low. A planned contrast testing the predicted interaction was significant, $t(156) = 2.68, p < .01, r = .16$. Furthermore, singular causation ratings were overall higher when α was low. A planned contrast testing this predicted main effect was also significant, $t(156) = 10.47, p < .001, r = .63$. There was no main effect of causal power, $t(156) < 1.00$.

The additional causal power ratings showed that participants’ causal power representations were not distorted by the different causal latencies. In the low-causal power condition, the mean ratings for the fast and slow tower were $M_{fast} = 5.18 (SD_{fast} = 1.09)$ and $M_{slow} = 5.30 (SD_{slow} = 1.16)$. In the high-causal power condition, the mean ratings were $M_{fast} = 7.55 (SD_{fast} = 1.61)$ and $M_{slow} = 7.69 (SD_{slow} = 1.97)$. A mixed ANOVA with “causal-power query” as within-subject factor yielded a main effect only for causal power, $F(1, 156) = 121.42, p < .001, \eta_G^2 = .39$, confirming that subjects in the high-causal power condition gave higher causal power ratings than subjects in the low-causal power condition. Importantly, there was neither a main effect of causal latency, $F(1, 156) < 1$, nor an interaction effect of causal latency and causal power, $F(1, 156) < 1$. We can thus rule out that the results were driven by different causal power representations in the different latency conditions.

General Discussion

To assess singular causation, more than just the causal powers of the potential causes need to be considered. Even when a cause factor is sufficient to generate the effect, it is still not the actual cause of the outcome if it was preempted by a competitor. To assess the probability of preemption, temporal information about the potential causes needs to be considered in combination with information about their power. We have discussed the roles of two relevant types of temporal

information: the difference of the onset times of the competing causes, and the latencies with which they generate their effects. To model the latencies of causes we used gamma distributions, which allows us to estimate the size of the α parameter in our model. The results of two experiments that we presented were explained well by our model, but not by Cheng and Novick’s (2005) power PC model of causal attribution which neglects temporal information.

The type of situation we considered here represents only a subset of possible cases. For example, unlike in our scenario reasoners often experience the latency between the target cause and outcome. Consider, for instance, a situation in which a person takes an aspirin and after twenty minutes gets relief from her headaches. In such cases, the experienced delays need to be used to estimate α . We plan to use dynamic test scenarios in future experiments to test such contexts.

Another noteworthy characteristic of our test scenario was that all potential causes of the effect were actually observed. In most real-world situations, however, reasoners observe only a subset of the potential causes. Other than in our test scenario, reasoners often are confronted with uncertainty concerning the presence of alternative causes. Although we have not considered these situations here, our model can be applied to them, too. For example, in situations in which only the target cause is observed but not alternative causes, information about the temporal distribution of the effect in the absence of the target cause needs to be considered to estimate α . This temporal distribution can be modeled with exponential functions (see, e.g., Bramley et al., in press). We plan to investigate such situations in future studies.

Acknowledgments This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Grants WA 621/24-1 and MA 6545/1-2, the latter as part of the priority program “New Frameworks of Rationality” (SPP 1516). We thank Louisa Reins for helpful discussions and assistance with the experiments.

References

- Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (in press). The role of time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*.
- Cartwright, N. (1989). *Nature’s capacities and their measurement*. Oxford: Clarendon Press.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.
- Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*, 694–706.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.
- Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*, 184–195.
- Paul, L. A., & Hall, E. J. (2013). *Causation: A user’s guide*. Oxford University Press.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
- Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*, 242–257.
- Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

Appendix G

Stephan, Tentori, Pighin, and Waldmann (in preparation)

Stephan, S., Tentori, K., Pighin, S., & Waldmann, M. R. (in preparation). Interpolating Causal Mechanisms: The Paradox of Knowing More.

Interpolating Causal Mechanisms: The Paradox of Knowing More

Simon Stephan¹, Katya Tentori², Stefania Pighin², & Michael R. Waldmann¹

¹University of Göttingen, Germany

²University of Trento, Italy

Author Note

Correspondence concerning this article should be addressed to Michael Waldmann, Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany. E-mail: michael.waldmann@bio.uni-goettingen.de. This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (WA 621/24-1).

Abstract

Causal knowledge is not static, it is constantly modified based on new evidence. The present set of five experiments explores one important case of causal belief revisions, causal interpolations. The prototypic case of an interpolation is a situation in which we initially have gathered knowledge about a direct causal or covariational relation between two variables but later become interested in the mechanism linking these two variables. Our key finding is that interpolations tend to be misrepresented, which leads to the paradox of knowing more. The more we know about the mechanism, the less predictable we seem to find the effects (i.e., weakening effect). In all experiments we found that despite identical learning data the predictive probability linking two variables C and E ($C \rightarrow E$) was rated reliably higher than the same predictive probability when a component of the mechanism (M) was interpolated (e.g., $C \rightarrow M \rightarrow E$). Our general theoretical explanation behind this stable weakening effect is that people have difficulties distinguishing between interpolations and lengthening scenarios (and other augmentations) in which causal variables are added on the effect side of the chain. Several experiments are presented that support specific predictions derived from this theory.

Keywords: causal reasoning, belief revision, probabilistic reasoning, causal Bayes nets

1. Introduction: Knowledge Revision in Everyday Causal Reasoning

Research focusing on causal reasoning generally explores how people acquire and use knowledge about relations between causes and effects. Most theories start with the assumption that the world can be categorized into a given set of variables that can be arranged in a causal network. Experiments in this field often focus on the question of how people learn and reason about this set of causal variables (see Waldmann, 2017, for overviews).

There is little research about the dynamic process of extending and deepening our knowledge (but see Bramley, Dayan, Griffiths, & Lagnado, 2017). We do not only learn about causal relations between given variables, we also acquire knowledge about new variables or acquire deeper knowledge about the mechanisms mediating the observed causal relations. Dynamic *causal belief revision* is a hallmark of scientific but also of everyday reasoning. For example, we may first learn that smoking leads to a disease, that a specific drug relieves our headache, or that pressing a switch turns a device on, but later we may become more curious and try to figure out how these causal contingencies are actually generated by underlying mechanisms. This process may involve the discovery of additional variables and the extension of a given causal network.

The process of re-representing our causal knowledge into more elaborate causal models can be described in various ways. A popular theoretical approach of how to represent causal knowledge are *causal Bayes nets*, which represent this knowledge as a set of variables linked by directed causal arrows (see Fig. 1 for an example) (see Rottman & Hastie, 2014; Rottman, 2017; Waldmann & Hagmayer, 2013, for reviews). The basic building block of causal Bayes nets are direct causal relations between a cause and an

effect, but these direct relations can be combined into indirect ones forming causal chains or more complex kinds of networks (see Fig. 1).

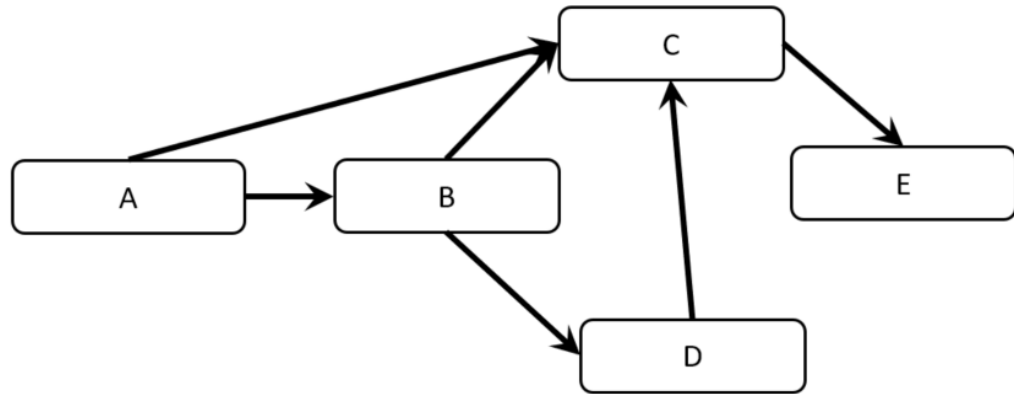


Figure 1. Example of a causal Bayes net depicting the causal relations between five variables, A, B, C, D, E.

In a given stage of our knowledge acquisition process, we may have acquired knowledge about a specific causal network. However, even when we are confident that our model is adequate and have collected data that support our assumptions about the structure of the model and the strengths of the causal relations, we may still want to elaborate the model later by adding variables.

It is important to note that the direct causal relations within a causal model are only direct relative to the chosen set of variables. Causal models are *frame-relative*, according to the analysis of Spohn (2012). It is always possible to turn a direct causal relation into an indirect one by interpolating new variables that mediate between the previously directly linked ones. Within causal Bayes net representations, the newly formed causal chains can then be considered representations of our current understanding of the mechanisms mediating the causal contingencies we already know from previous observations. In the present research we will study how people reason in situations of causal belief revision. We

will mainly focus on causal chains, but later we will demonstrate that our findings can be generalized to more complex structures.

1.1 *Revising Beliefs about Causal Chains*

There are various ways in which knowledge about a causal chain can later be revised: First, it could turn out that the causal directions within a chain do not adequately represent the causal situation. Thus, one possibility for revision concerns the *structure* of the causal model. Second, given that causal strength is estimated based on a limited set of data it may later be discovered that the strength estimates are distorted or that the researcher failed to control for relevant confounds when estimating strength. Third, new relevant variables connected to the known network could have been discovered in the meantime, which lead to an augmented network. In the case of causal chains, adding variables may lead to *lengthening* of the chain. Fourth, mechanisms mediating between variables could later be discovered, which leads to the *interpolation* of further variables.

As an example for an interpolation, our knowledge acquisition process may start with observations that bolster our belief in a stable causal contingency between smoking and lung cancer, for example. Therefore, in an initial causal model representation smoking would play the role of a direct cause of lung cancer. Later we may become interested in how this relation is mediated. This may lead to a more elaborate physiological representation turning the direct into an indirect causal relation (e.g., smoking → genetic alterations → lung cancer). The key question of the present research concerns the question of how extensions of causal knowledge about chains affect our beliefs in the statistical relations between causal variables. We will in the introduction focus on the two cases of extending causal chains, *lengthening* and *interpolating*. Our experiments will then focus on interpolations, which have not yet been studied in detail.

1.1.1. Lengthening of Causal Chains

Lengthening a given chain by adding variables at the beginning or end of the chain typically leads to a weakening of probabilistic relations between the given and new variables proportional to the length of the distance between the variables within the newly discovered chain. A key assumption here is that the initial chain stays invariant, additional variables are just added at one of the outer sides. Imagine, for example, that a direct causal relation had been discovered between the variable JPH3 (a fictitious mutation of a gene) and Hepatocytosis (a fictitious disease) (see Fig. 2A). Since other known and unknown outside variables typically additionally affect causal relations, the causal relation between JPH3 and Hepatocytosis will most likely be probabilistic on the observational level.

There are many ways to measure causal strength, but a standard method to assess the strength of the causal relation is by using the contingency measure ΔP (see, for example, Perales, Catena, Candida, & Maldonado, 2017). ΔP equals the difference between $P(e|c) - P(e|\neg c)$ (with e representing the effect, c the target cause, and $\neg c$ the absence of the cause). ΔP can range between -1 (for a perfectly inhibitory relation) and +1 (for a deterministic generative relation).

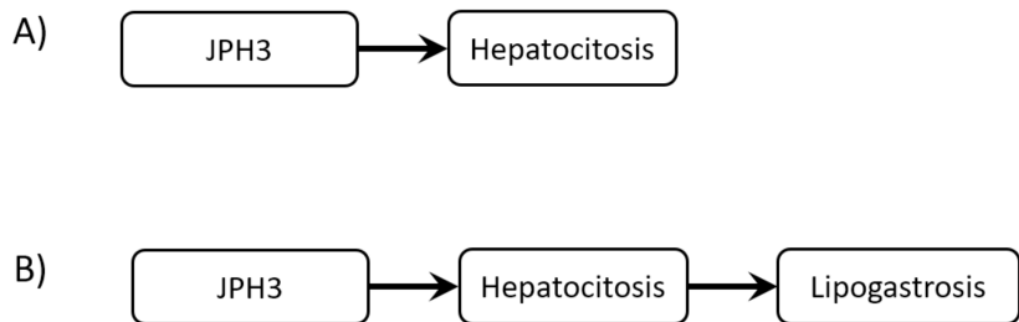


Figure 2. Illustration of the lengthening of an initial causal chain (A) into a causal chain with an additional effect (B).

Figure 2B shows an example of a lengthening of the chain shown in Figure 2A. The chain contains now three variables. It is assumed that it has been discovered that Hepatocytosis directly causes another disease, Lipogastrosis. Assuming that the Markov condition holds, which in this case means that the new probabilistic relation between Hepatocytosis and Lipogastrosis is independent of whether JPH3 is present or absent (see Mayrhofer & Waldmann, 2015), the following relation holds:

$$\Delta P_{JPH3-Lipo} = \Delta P_{JPH3-Hepa} \times \Delta P_{Hepa-Lipo} \quad (\text{Equation 1}).$$

The equation entails that the contingency between JPH3 and Lipogastrosis is only equal to the two contingencies between the two directly linked variables (i.e., JPH3-Hepatocytosis, Hepatocytosis-Lipogastrosis) when these relations are deterministic, otherwise a *weakening* effect is expected. The further the distance between variables (i.e., the longer the chain), the weaker the contingency between JPH3 and the final effect is expected to be. It is important to note that this effect is based on the assumption that variables are added to otherwise invariant chains. The exact functional form of the probabilistic relations between the three variables will of course be different when the Markov condition does not hold. But since our focus will not be on lengthening but rather interpolations, we will not pursue the role of this constraint further.

Psychological research has shown that reasoning about causal chains is consistent with this multiplicative lengthening constraint (Equation 1) (Ahn & Dennis, 2000; Baetu & Baker, 2009; Jara, Vila, & Maldonado, 2006). Presented with individual pairwise causal links that are later combined into a three-variable chain, subjects tend to believe that the initial cause indirectly causes the final effect, thus demonstrating a belief in causal transitivity. Moreover, their judgments are generally consistent with the multiplication constraint expressed in Equation 1. Interestingly, reasoning about chains even tends to express a transitivity bias

when the presented data actually violate the Markov condition and therefore are inconsistent with transitive chains (v. Sydow, Hagmayer, Meder, & Waldmann, 2010; v. Sydow, Hagmayer, & Meder, 2016). There seems to be a tendency to assume the validity of the Markov constraint even when it is violated, at least in these tasks.

Another study testing whether causal reasoning is consistent with the normative predictions of Bayes nets was conducted by Bes, Sloman, Lucas, and Raufaste (2012). The study investigated a number of causal models and showed judgment biases that can be better explained by a theory of explanatory ease than by Bayes nets. However, some of the results seem consistent with Bayes nets. In their Experiment 2, Bes et al. presented three causal variables that were sufficiently neutral so that they could be arranged in different causal models through verbal instructions. Moreover, some vague information was given about the strength of the covariation between the three variables suggesting that for all variable pairs 40% have both high values, another 40% have both low values, and 20% have mixed values. After this information, causal model instructions were provided. For our project, the chain conditions are the most relevant ones. In one condition, the *predictive direct chain condition*, A was the direct cause of B and B of C ($A \rightarrow B \rightarrow C$), whereas in one of the other conditions, the *indirect predictive chain condition*, B was indirectly caused by A, $A \rightarrow C \rightarrow B$. As a test question measuring probabilistic intuitions, subjects were requested to respond to rate predictive conditional probabilities of effects conditional on the presence of a cause using a rating scale ranging from 0% to 100% (e.g., $P(B|A)$).

The key finding was that despite identical data the conditional probability of B given A was rated significantly higher when the instructed causal chain model linked them directly (direct predictive causal chain) than when they were indirectly linked (indirect predictive causal chain).

This finding was replicated in a follow-up study (Experiment 3) in which, prior to causal model instructions, trial-by-trial data were presented showing individual cases. In a *long learning condition*, 60 trials were presented. In 24 cases all variables had high values, in another 24 they were all low, the rest of the patterns (with mixed values) were presented twice each. In a *short learning condition*, only 5 cases were presented with 2 all being high and 2 all being low (and one additional random mixed case). Again, the indirect causal chain yielded lower ratings for the probabilistic relation between A and B than the direct causal chain. This effect was not sensitive to the length of the training phase.

In sum, in both experiments Bes et al. (2012) presented evidence consistent with a lengthening effect. Given the symmetry of the statistical relations between the three variables, this study provides further evidence in the belief consistent with a transitivity assumption although the data violate the Markov condition so that they are actually ambiguous with respect of the underlying causal model. These results along with the studies by v. Sydow and colleagues (2010, 2016) seem to indicate that people focus on the structure of causal models rather than data patterns (e.g., sample size; Markov condition).

1.1.2. *Interpolation in Causal Chains*

In the previous section we have shown that in lengthening scenarios subjects tend to expect a weakening of probabilistic dependencies with increased distance within the causal chain. People even assume weakening with longer chains when the formal preconditions (i.e., Markov condition) do not hold.

Our main focus in the present research will be another type of extension of chains that is generated by interpolations of mediating variables. Interpolations are frequent in contexts of causal discovery of mediating mechanisms. Reversing the revision process in the example from the last section, we may, for example, first have obtained reliable covariation

knowledge indicating that JPH3 causes Lipogastrosis. In a causal model we would represent this as a direct causal relation between the two variables (see Fig. 3A), which can be used for predictions, diagnoses, or causal interventions. However, later we may want to know *how* JPH3 exerts its influence on Lipogastrosis. We may then learn that the causal relation is mediated within a causal chain by Hepatocytosis. Ordinarily we see again a three-step chain, like the one in Figure 2B, that appears longer after the interpolation, but this effect is due to a very different process than the lengthening we obtain when adding variables at the beginning or end of a chain. Here, a previously direct causal relation is turned into an indirect one by zooming in on the causal relation and adding an intermediate step. Our key question is whether interpolations lead to a similar weakening effect as cases where a given chain is lengthened at the outer sides of the chains. This is a novel question that has not been addressed in the literature so far.

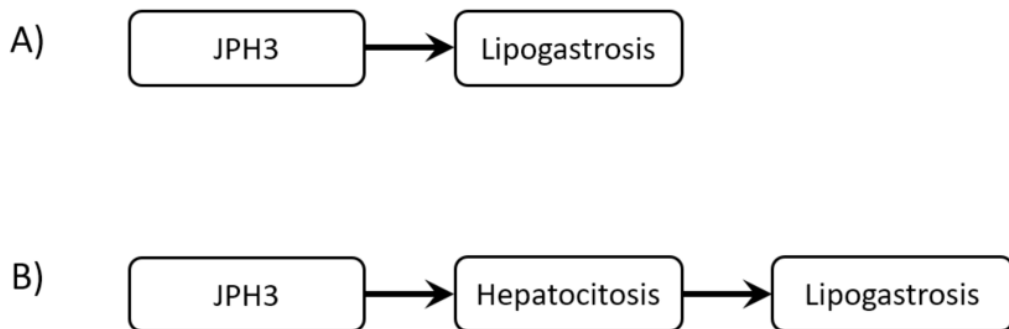


Figure 3. Example for the lengthening of causal chains resulting from the interpolation of variables.

1.2. The Paradox of Knowing More

To preview the results of our research presented later, we have observed a stable weakening effect also in causal interpolations. For example, once reasoners are informed that the causal relation between JPH3 and Lipogastrosis is in fact mediated by the variable Hepatocytosis, they tend to assume that the previously observed probabilistic dependency

between JPH3 and Lipogastrosis should now have become weaker. Interestingly, we have discussed such interpolation scenarios with colleagues from psychology, philosophy, and computer science, and many had the initial reaction that such a weakening is perfectly rational. We actually did initially not agree about this issue among ourselves. Later, upon reflection, most people often changed their minds, though. We found these intuitions by experts interesting and wanted to explore how they can be explained.

1.3. *Is A Weakening Effect Rational in Causal Interpolations?*

Although our focus will be on a psychological explanation of the weakening effect in interpolations, it is an interesting question whether weakening is a fallacy or can be defended as a rational response. In the past years, a number of phenomena that initially appeared to reflect irrational biases were, more or less convincingly, re-interpreted as results of rational processes (e.g., Costello & Watts, 2014; Juslin, Nilsson, & Winman, 2009; Hertwig & Pleskac, 2010; Kareev, 2000). Can the weakening effect similarly be explained as a rational response? One can address this question from a metaphysical or an epistemic point of view, which we will discuss in turn.

Metaphysical Analysis. From a metaphysical point of view the answer seems clear. As long as we stay in the same context as the learning situation, interpolations of variables should not change the contingencies between the observed variables. One popular philosophical approach to represent causal relations is by assuming that the causal events are connected by a spatio-temporally continuous causal line on which some kind of conserved quantity (e.g., momentum) (e.g., Dowe, 2000) is transferred in an active causal relation. Within this framework, measured causal variables represent abstracted events on this causal line. It seems self-evident that further measurements along this line should not change the relationship between the initially chosen cause and effect variables.

Causal lines cannot only represent deterministic relations, for which no lengthening effect is predicted anyway. They can also represent more complex models with additional unobserved lines impinging on the observed one. In the Bayes net literature, it has been shown that probabilistic relations can be re-represented as deterministic chains with hidden enablers or preventers (i.e., quasi-deterministic networks) without loss (see Spirtes, Glymour, & Scheines, 2000; Pearl, 2000). Measuring additional variables on a chain embedded in a quasi-deterministic network should also not change the initially represented causal relation, according to our intuition.

Another way to demonstrate that interpolations should not change contingencies can be seen in Figure 4, which represents a mechanism as a Bayes net representation of a causal chain. Figure 4A shows the representation of the direct causal relation between JPH3 and Lipogastrosis at some time point, t1. Now assume, we later discover at t2 that Hepatocytosis mediates this relationship (see Fig. 4B). Where was Hepatocytosis at t1? Again, a natural assumption is that Hepatocytosis already mediated the relation between JPH3 and Lipogastrosis at t1, this mediation relation was just unknown to us at t1. Discovering it at t2 should not change anything, as long as we do not change into a different context with different variables affecting the covariation (see General Discussion). In fact, if we assume that in the unknown underlying Bayes net (let's call it God's Bayes Net), there is an infinite number of mediating variables between JPH3 and Lipogastrosis that await discovery (Fig. 4B shows a fragment), then further discoveries (e.g., Fig. 4C) should also not change the strength of the relation between the initially discovered two variables.

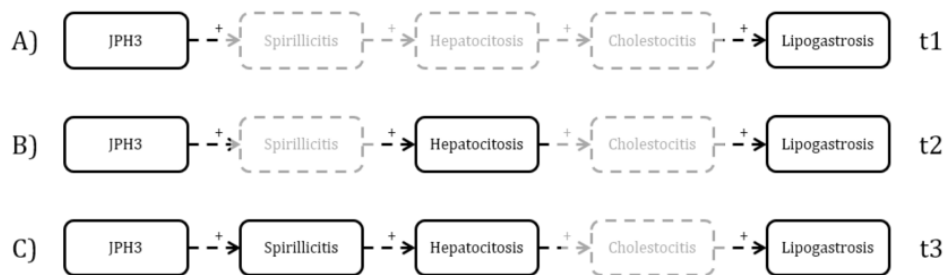


Figure 4. Illustration of the process of causal mechanism discovery within God's Bayes Net.

Epistemic Analysis. Another possible path to justify the weakening effect might focus on the fact that the initial observations are typically unreliable. Contingency or other kinds of probability estimates rely on samples so that a specific amount of uncertainty is always attached to these estimates. In fact, many of the attempts mentioned above to explain apparent biases and fallacies as rational argue that they are rational responses to uncertainty.

We do not believe that the weakening effect in causal interpolations can be explained as a rational response to uncertainty in our experiments, however. For one, in our experiments we present a sufficiently large sample that should allow for fairly reliable estimates. Moreover, we were using causal strength parameters that were sufficiently far away from the extremes 1 and 0, so that no ceiling or floor effects are to be expected. If uncertainty is construed as unreliability of the measurement, a symmetric confidence interval is expected which should lead to a belief in a range of possible true values which include lower as well as higher values.

Uncertainty could also be construed as a tendency to incorporate a prior in the estimate. However, in previous causal research strength priors have been postulated that tend toward sufficiency or necessity rather than lower values than the ones we used, which would

actually predict more extreme estimates (see Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Mayrhofer & Waldmann, 2016; Yeung & Griffiths, 2011, 2015).

A weakening effect would on this account only be predicted if the strength prior refers to individual direct links regardless of whether the causal chain is the product of lengthening or interpolation. But collapsing these two processes clearly misrepresents these two very different situations and is therefore not normative. In fact, we later present a related model as a psychological explanation of the weakening effect in causal interpolations (see sections 1.4, 4, 5).

Finally, and most importantly, our key manipulation is whether we inform subjects after they have initially observed a contingency between two variables that researchers had discovered a mediating variable or whether the researchers believed that the two covarying variables are directly linked. In none of the experiments we presented any new data about the causal variables, which could disconfirm the initial estimate. Even if subjects attached uncertainty to the initial estimate, and had for some reason a tendency to provide more conservative estimates for what they have seen, there is no normative statistical reason to revise the probability estimates differently in the two conditions (direct vs. indirect chain), as long as they are the result of interpolation. A possible uncertainty about the initial estimate should not be changed by verbal instructions about the underlying mediating causal model.

The Paradox of Knowing More. Another argument supporting the claim that a weakening effect may not be rational in causal interpolations, refers to an interesting counterintuitive implication of a generalization of this effect, which we express by using the label “paradox of knowing more.” Under the assumption of a rational weakening effect, the longer the chain becomes, that is, the more mediating steps we discover about the causal relation, the weaker the dependencies would become. This does not sound sensible. The better we

understand causal mechanisms in the world, the better we should be able to make predictions, diagnoses, or interventions. An increase of knowledge should not lead to increased unpredictability of the world.

1.4. *Psychological Accounts of the Weakening Effect*

Regardless of whether a weakening effect is considered rational or not in causal interpolations, the question still remains how a psychological theory could explain it. Thus, the main focus of our research is testing psychological factors that might underlie the weakening effect in causal interpolations. We will here only briefly preview our hypotheses and the experiments; more detailed accounts will be given in the introductions of the studies.

The experimental series will start with three studies in which we set the stage for later studies by demonstrating the existence of the weakening effect (Experiments 1, 2). We will test the effect using various types of variables. Next, we will study possible psychological factors underlying the weakening effect. The factors we test are not mutually exclusive, they are interdependent.

Explicit vs. Implicit Representation of Enablers and Disablers. The first hypothesis assumes that causal interpolations lead to longer chains with more variables, which potentially highlights multiple possibilities of how things can “go wrong.” Subjects might be led to consider enablers of mediating variables which might be missing, disablers which might be present, or alternative causes. If, for example, we just represent smoking and lung disease we might consider factors additionally influencing lung disease. If we represent the relation mediated by genetic alterations, for example, then we might consider additional factors influencing genetic alterations. Of course, normatively these additional factors do not arise simply because we have discovered the relevance of the mediating factor, for example

genetic alterations. They should already be causally effective before we made this discovery, and be reflected in the initial simpler causal model, at least implicitly. But psychologically it may make a difference whether we explicitly represent the mediating variables along with potential enablers, disablers, and alternative causes, or whether these moderating variables only implicitly affect the observed causal relationship. We will focus on disablers as an example for our hypothesis and test it by manipulating whether disablers are explicitly mentioned or not (i.e., explicit vs. implicit representations) (Experiment 3).

Confusing Lengthening and Interpolations. A second hypothesis we are going to test assumes that people tend to be sensitive to the length of chains but do not clearly differentiate between a situation in which a pre-existing chain was lengthened and one in which a chain was extended by interpolating additional variables. On the surface, a chain with several variables looks the same, regardless of its history (see Figure 5).

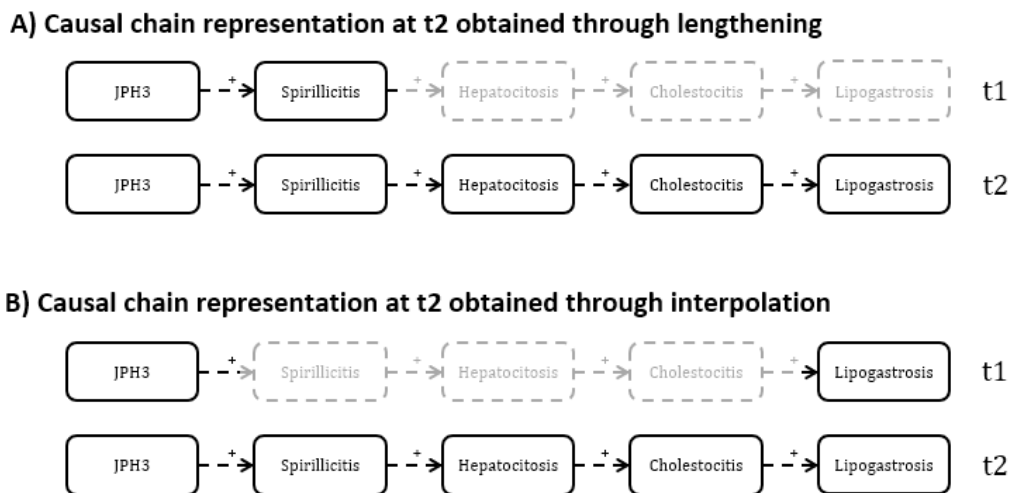


Figure 5. Illustration of how lengthening and interpolation might lead to the same representation. A) shows a causal chain at t2 following a discovery of new factors downstream of the initial effect. B) shows the same causal chain at t2 but this time obtained through the interpolation of variables between the initial cause and final effect.

Since in lengthening scenarios probabilistic relations normatively tend to weaken, we may assume the same for chains of the same length that are generated by interpolations. One prediction that this hypothesis entails is that reasoners should be expected to assume that the contingencies linking known adjacent events in a chain stay invariant regardless of how long the chain is. The causal strengths of the links of the chain prior to lengthening do not change when variables are added. Of course, further links might be added with stronger or weaker causal strengths but their strengths are again independent of the final length of the chain in lengthening scenarios.

By contrast, when new variables are interpolated, the causal contingencies connecting the directly linked variables should systematically become stronger. This is a direct consequence of Equation 1, the product rule. If, for example, $\Delta P_{JPH3-Lipo}$ equals 0.7, then the causal strengths of the mediating links should become on average stronger, the more variables are interpolated (see section 5). Thus, if interpolations are misrepresented as lengthening, causal strength estimates for the individual links should be observed that stay invariant independent of the length of the chain. If they stay invariant, then their product would deviate further and further away from 0.7 the longer the chain becomes, which would lead to a weakening effect. This hypothesis will be tested in Experiment 4.

Interpolating Complex Networks. Our main focus in the present research will be on causal chains. Here we have a clear contrast between a normative weakening effect observed in lengthening scenarios, and a non-normative weakening effect observed in interpolations. How about interpolating more complex network structures between two variables (cf. Figure 1)? A specific target relation (e.g., JPH3-Lipogastrostrosis) might turn out to be mediated by a complex network with multiple cause-effect relations between the two initial variables. Here we also have an interesting asymmetry.

In situations in which an existing causal relation is extended by discovering a network of variables surrounding this causal relation (corresponding to a lengthening of chains), predictions about probabilistic relations depend on the structure and parameters linking the network variables. By contrast, the strength of the initially observed causal relation should again stay invariant if a network is interpolated between the two variables. Regardless of the complexity and the parameters connecting the mediating variables, the causal contingency should not systematically change, for the same reasons why we do not expect such a change in simple chains. However, similar to our findings with causal chains, we expect that causal properties of the interpolated network will systematically cause revisions of the estimates concerning the initially observed two variables. This hypothesis will be tested in Experiment 5.

2. Experiment 1

The goal of Experiment 1 was to set the stage for the project by testing whether we will find a weakening effect after causal interpolations. We generally are using fairly abstract materials to avoid distortions by prior knowledge. Therefore, the causal variables refer to fictitious unknown entities. Experiment 1 tests a paradigm we are using in all other studies. In Phase 1, subjects are presented with trial-by-trial learning information about two covarying variables. This learning phase allows subjects to acquire knowledge about the degree of covariation between the two variables suggesting a causal link. Subsequently, subjects are either told that scientists had discovered that the two variables are directly causally related or that scientists had discovered that the two observed variables are in fact indirectly related, with a newly discovered variable mediating the two variables on a causal chain. No new data are shown in Phase 2. Then the final test question requests subjects to estimate the conditional probability of the effect variable given the cause variable referring

to the two variables shown in the initial learning phase (see Bes et al., 2012, who have used a similar test question). If we observed a weakening effect, the conditional probability estimate is expected to be lower in the indirect than the direct condition, despite identical learning data in both conditions.

2.1 Method

2.1.1. Participants

One hundred and forty subjects (100 female, one subject indicated to be neither male nor female, $M_{Age} = 36.69$, $SD_{Age} = 12.50$) were recruited via the online platform Prolific (www.prolific.ac). This sample size allows it to detect a medium effect size of $d = 0.50$ with more than eighty percent probability. All subjects were native English speakers and had at least an A-level degree. They were paid £ 0.70 for their participation.

2.1.2 Design, Materials, and Procedure

As cover story we used a fictitious scenario according to which biologists were interested in studying the statistical relation between the mutation of the gene *JPH3* (J) and the gastrointestinal disease *Lipogastrosis* (L), which is characterized by an excessive accumulation of fats in the digestive tract. Subjects were further instructed that the biologists had conducted a study in which they examined two random samples of mice, one consisting of mice carrying the mutation and a second one not carrying the mutation. We informed subjects that the results of the biologists' study will be presented in serial order and that their task will be to examine the results thoroughly and not take any notes during the study. Before subjects could proceed to the learning task, they had to pass an instruction check involving three questions.

The contingency that we presented to subjects in the subsequent learning task is shown in Table 1. The probability of Lipogastrosis given a mutation of JPH3 was $P(L|J) = 0.75$ and

the base rate of Lipogastrosis in the absence of Lipogastrosis was $P(L|\neg J) = 0.21$. Hence, the contingency was $\Delta P = 0.54$.

Table 1

Contingency presented to subjects in Experiment 1

	L	\neg L	$P(L J)$	$P(L \neg J)$	ΔP
J	18	6	.75	.21	.54
\neg J	5	19			

As learning task, we used a classic trial-by-trial observational learning task in which the forty-eight cases summarized in Table 1 were presented to participants on a computer screen in random order. The learning phase had to be initiated by a click on an “Examine Results” button presented at the bottom of the screen. An example of what the screen looked like during the learning phase is shown in Figure 6. The presence of either factor (J or L) was indicated by a yellow text box, while the absence of a factor (\neg J or \neg L) was indicated by a gray text box. The case numbers were displayed in the upper left corner of the screen. Each case was displayed for four seconds followed by a white mask displayed for 500 ms. The duration of the learning task was roughly three minutes.

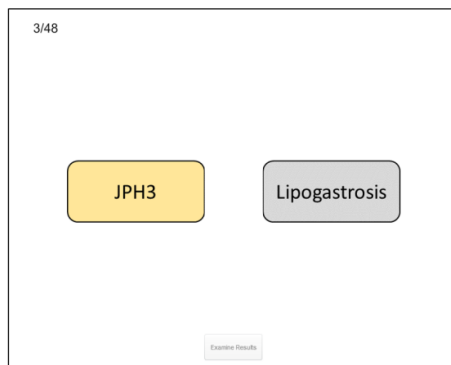


Figure 6. Illustration of the trial by trial learning task used in Experiment 1.

After subjects had finished the learning task, they were given information about the causal model assumed to underlie the observed relation between JPH3 and Lipogastrosis. Depending on condition, subjects were either instructed that JPH3 and Lipogastrosis were *directly* causally related or that they were *indirectly* causally related in a chain containing one interpolated variable between JPH3 and Lipogastrosis. In the *direct causal relationship* condition, subjects were presented the following text along with the illustration shown in Figure 7.

*“Please read the following new information:
The biologists later found out that JPH3 and Lipogastrosis are in fact directly causally related as illustrated in the figure below. That is, the JPH3 mutation can sometimes lead to Lipogastrosis. This is indicated by the arrow (with a + sign) that goes from JPH3 to Lipogastrosis. Other factors can also influence the disease.”*

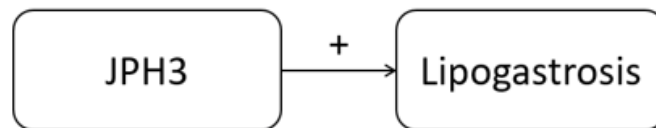


Figure 7. Illustration of the causal model shown to subjects in the *direct causal relationship condition* of Experiment 1.

Subjects in the *indirect causal relationship* condition were presented the following text together with the illustration shown in Figure 8.

*“Please read the following new information:
The biologists later found out that JPH3 and Lipogastrosis are in fact indirectly causally related as illustrated in the figure below. Specifically, the JPH3 mutation can sometimes lead to Hepatocytosis, an abnormal occurrence of hepatic enzymes. This is indicated by the arrow (with a + sign) that goes from JPH3 to Hepatocytosis. Finally, Hepatocytosis can sometimes lead to Lipogastrosis. This is indicated by the arrow (with a + sign) that goes from Hepatocytosis to Lipogastrosis. Other factors can also influence the disease.”*



Figure 8. Illustration of the causal model shown to subjects in the *indirect causal relationship* condition of Experiment 1. The *interpolated* variable was either *Hepatocytosis*, *Cholestocytis*, *Spirillicytis*, or *Paracelocytis*.

We used different labels for the interpolated variable of the chain. Subjects either learned that the interpolated variable was *Hepatocytosis* (see Fig. 8), described as an abnormal increase in hepatic enzymes, or *Cholestocytis*, described as an abnormal occurrence of cholesterol, or *Spirillicytis*, described as an infection colonizing the gut, or *Paracelocytis*, described as a dysfunction of the process involved in fat metabolism.

After the participants had read the causal model information, they proceeded to the test screen. Subjects read that the biologists were now inspecting a new mouse that they had randomly sampled and of which they had noticed that it carries the mutation. Subjects in the *direct causal relationship* condition were shown the illustration depicted in Figure 9.

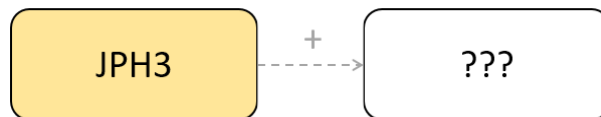


Figure 9. Illustration shown during the inference phase in the *direct causal relationship* condition of Experiment 1.

Subjects in the *indirect causal relationship* condition were shown the illustration depicted in Figure 10, with the interpolated variable being the one presented during the structure information phase.

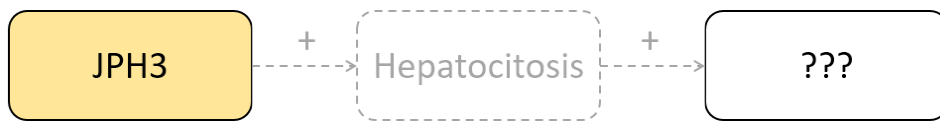


Figure 10. Illustration shown during the inference phase in the *indirect causal relationship* condition of Experiment 1. The *interpolated* variable was either *Hepatocytosis*, *Cholestocytis*, *Spirillicitis*, or *Paracelocytis*.

Subjects were then asked to estimate the predictive probability of Lipogastrosis given JPH3, $P(L|J)$. The phrasing of the test question, which referred to the image shown in Figures 9 and 10 was: “What do you think is the probability that the mouse also has Lipogastrosis?” Subjects provided their ratings on a slider ranging from 0 to 100 with the endpoints labeled “It is certain that this mouse does not have Lipogastrosis” and “It is certain that this mouse has Lipogastrosis.”

2.2 Results and Discussion

The results are summarized in Figure 11. As can be seen, subjects in the *direct causal relationship* condition gave ratings ($M = 68.47$, $SD = 16.70$) that were close to the normative value of $P(L|J) = 0.75$. As can also be seen, the ratings in the *indirect causal relationship* condition ($M = 56.30$, $SD = 21.58$) were lower than those of the *direct causal relationship* condition, as predicted. An independent t-test confirmed that the observed difference was significant, $t(138) = 3.73$, $p < .001$, $d = 0.63$.

The results of this experiment indicate that interpolating causal variables between two covarying causal variables changes reasoners’ representation of the observed probabilistic relation. We observed a “weakening effect” in the indirect compared to the direct causal relation although the interpolated variable was introduced as a mediator between the two variables rather than representing an additional variable in a lengthening scenario.

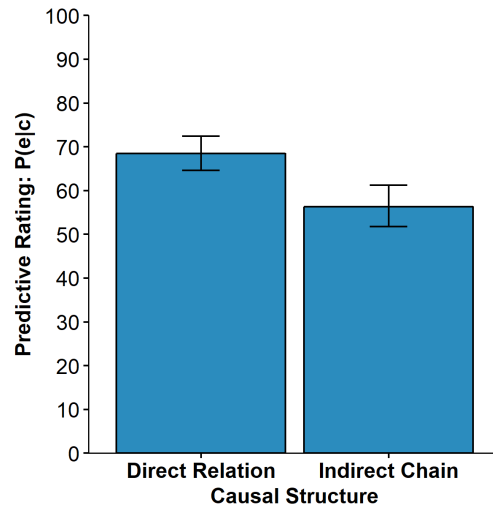


Figure 11. Results (means and 95% CIs) of Experiment 1.

3. Experiment 2

A characteristic of Experiment 1 was that the interpolated variables and the effect variable were always described as physiological: they were described as a particular symptom or a particular disease involving dysfunctional physiological processes. One could therefore argue that our results may be restricted to this kind of mechanism. To increase generality, the goal of Experiment 2 was to test if the weakening effect interacts with the type of variables that are being used.

We constructed three different versions of the cover story used in Experiment 1, varying the level of description on which the variables following the root cause are

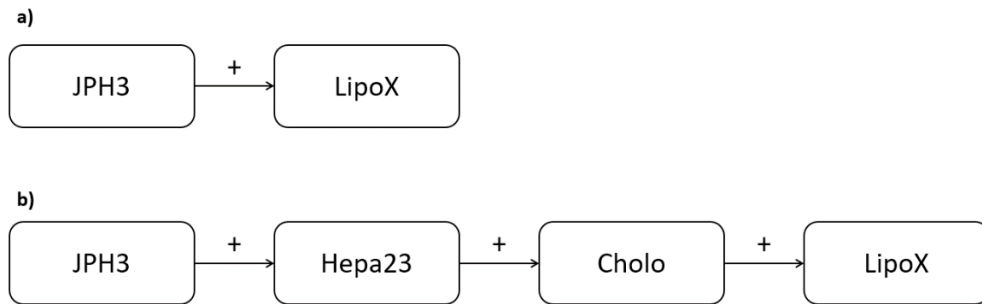


Figure 12. Causal graphs presented to subjects in Experiment 2 after the learning phase. In the direct causal relationship condition, subjects were presented the structure shown in a). In the indirect causal relationship condition, subjects were presented the structure shown in b).

described. We contrasted the physiological level with a genetic and a molecular level.

Furthermore, we introduced two interpolated variables in the indirect causal relationship condition instead of one.

3.1 Method

3.1.1. Participants

Five hundred and ten subjects (242 female, 265 male, three subjects indicated neither male nor female, $M_{Age} = 33.29$, $SD_{Age} = 14.73$) were recruited via the online platform *Prolific* (www.prolific.ac). This sample size allows it to detect small main effects and interactions of $d = 0.3$ with more than eighty percent probability. We planned for a small effect this time because we were uncertain whether our new level of description variable might influence the results. All subjects were native English speakers and had at least an A-level degree. They were paid £ 0.70 for their participation.

3.1.2. Design, Materials, and Procedure

The design was a 2 (causal model: direct vs. indirect causal relationship) \times 3 (level of description: physiological vs. genetic vs. molecular) between-subjects design. The cover story and the procedure were identical to the ones in Experiment 1, with the exception that, depending on condition, the interpolated variables (in the indirect conditions) and the effect

variable (in all conditions) were described either as a physiological, genetic, or a molecular-level process. The root cause was always the genetic mutation of JPH3 (as in Experiment 1).

Like in Experiment 1, subjects first read about a group of biologists who were interested in investigating the statistical relation between a mutation of the JPH3 gene and a particular disease. To be able to manipulate the type of description, we labeled the disease LipoX instead of Lipogastrosis. Depending on condition, the disease was either described as being “an abnormal modification of fats in the digestive tract” (physiological-level condition) or as being “an abnormal mutation of a gene in the digestive tract” (genetic-level condition) or as being “an abnormal change of the molecular structure of enzymes in the digestive tract” (molecular-level condition). Like in Experiment 1, subjects were then presented with the serial trial-by-trial learning task that conveyed the contingency between the JPH3 mutation and LipoX, and thereafter were presented with the causal model information. In the direct causal relationship condition, subjects were shown the illustration in Figure 12a, which was accompanied by the following instruction:

“Later, the biologists found out that JPH3 and LipoX are directly causally related, which is illustrated in the figure below. That is, the JPH3 mutation can sometimes lead to LipoX (indicated by the arrow with the plus sign that goes from JPH3 to LipoX), an abnormal modification of fat in the digestive tract [vs. an abnormal mutation of a gene in the digestive tract vs. an abnormal change of the molecular structure of enzymes in the digestive tract.”

The last part was varied according to condition. Note that the different variables within each chain mentioned the same type of abnormality (abnormal modification of fat, abnormal gene modification, or a change of the molecular structure of enzymes). Our goal here was to increase the semantic coherence of the mechanism so that we can rule out that we get weakening effects only when we interpolate mechanisms coming from different categories (e.g., different disease).

Table 2

Descriptions of the interpolated variables in the indirect causal relationship condition of Experiment 3.

Condition	Description
physiological	"Hepa23 (indicated by the arrow that goes from JPH3 to Hepa23 and the plus sign above the arrow), which is an abnormal modification of fat in the liver. Further, Hepa23 can sometimes lead to Cholo, an abnormal modification of fat of the spleen. Finally, Cholo can sometimes lead LipoX, an abnormal modification of fat in the digestive tract."
genetic	"Hepa23 (indicated by the arrow that goes from JPH3 to Hepa23 and the plus sign above the arrow), which is an abnormal gene mutation in the liver. Further, Hepa23 can sometimes lead to Cholo, an abnormal gene mutation in the spleen. Finally, Cholo can sometimes lead to LipoX, an abnormal mutation of a gene in the digestive tract."
molecular	"Hepa23 (indicated by the arrow that goes from JPH3 to Hepa23 and the plus sign above the arrow), which is a change of the molecular structure of enzymes in the liver. Further, Hepa23 can sometimes lead to Cholo, a change of the molecular structure of enzymes in the spleen. Finally, Cholo can sometimes lead to LipoX, an abnormal change of the molecular structure of enzymes in the digestive tract."



Figure 13. Illustrations shown in the test phase of Experiment 2. Subjects in the direct causal relationship condition were presented the structure shown a). Subjects in the indirect causal relationship condition were presented the structure shown in b).

In the indirect causal relationship condition subjects were shown the causal model

depicted in Figure 12b, accompanied by the following instruction:

“Later, the biologists found out that JPH3 and LipoX are indirectly causally related by a chain that is illustrated in the figure below. Specifically, they found out that the JPH3 mutation can sometimes lead to [...], where the information in the square brackets varied according to condition.”

The instructions that were given in the different conditions are listed in Table 2.

After subjects read the causal model instructions, they proceeded to the test scenario. Like in Experiment 1, subjects read that the biologists were now inspecting a new mouse that they had randomly sampled and which carries the mutation. Subjects in the *direct causal relationship* condition were shown the illustration depicted in Figure 13a. Subjects in the *indirect causal relationship* condition (causal chain) were shown the illustration depicted in Figure 13b. As in Experiment 1, subjects were asked to estimate the predictive probability of LipoX given JPH3, $P(L|J)$. The phrasing of the test question was: “What do you think is the probability that the mouse also has LipoX?” Subjects provided their ratings on a slider ranging from 0 to 100 with the endpoints labeled “It is certain that this mouse does not have LipoX” and “It is certain that this mouse has LipoX.”

3.2. Results and Discussion

The results are summarized in in Table 3. As can be seen, subjects tended to give lower ratings in the indirect causal relationship conditions than in the direct causal relationship conditions, irrespective of the level of description and type of variable. A 2 (causal model: direct vs. indirect) \times 3 (level of description: physiological vs. genetic vs. molecular) factorial ANOVA revealed that the main effect of “causal model” was significant, $F(1, 504) = 20.78, p < .001, d = .41$, confirming again a weakening effect. There was, by contrast, no effect of “level of description”, $F(2, 504) = 1.10$, and also no interaction between “causal model” and “level of description”, $F(2, 504) = 0.37$.

Table 3

Summary of the results in Experiment 2.

	Direct causal relationship			Indirect causal relationship		
	Physiological level	Genetic level	Molecular level	Physiological level	Genetic level	Molecular level
Mean	68.66	68.25	66.28	59.18	62.34	58.06
<i>SD</i>	16.63	17.92	16.47	21.15	21.99	21.88
Median	71.00	71.00	68.00	61.00	67.00	63.00
95% <i>CI</i>	3.59	3.87	3.55	4.56	4.74	4.72

This experiment therefore provides further evidence that the weakening effect is robust and that it is not dependent on the level of description and the degree of semantic coherence of the instructed causal mechanism.

4. Experiment 3

The previous experiments aimed to establish that, given a specific contingency between an invariant cause and effect, C and E, reasoners’ predictive probability estimations $P(E|C)$ are lower when they believe that C and E are indirectly connected by a causal chain

than when they believe that C and E are directly causally connected. We showed that this weakening effect in causal interpolations occurs robustly across different contexts and tasks.

The primary goal of Experiment 3 was to investigate a first possible psychological explanation of the weakening effect. One explanation for the weakening effect is that reasoners might perceive indirect causal relationships to be more prone to “failure” than direct ones. Whereas a direct causal relationship may appear stable, an indirect relationship may lead people to reflect about possible ways the mediating variables may be affected by outside variables, such as disablers, enablers, or alternative causes. The more variables the chain contains, the more outside variables attached to the different variables on the chain may be considered. Since the presence of disablers or the absence of enablers may disrupt the causal flow, a lowering of predictive probability ratings might be observed. By contrast, in a direct causal relation there are only two events that can be affected by outside variables.

Normatively there should be no difference between explicit and implicit presentations of variables in cases of interpolation. In the initially represented direct causal relationship these disrupting possibilities are also reflected in the probabilistic relations inherent in the observed non-deterministic contingency. In direct relations, the existence of disablers is implicitly implied by a lowered predictive probability between C and E in the data because something must have been the cause for the effect being absent when the cause is present. When this direct relation is subdivided into a chain, potential disruptions should just be subdivided along the chain, by assuming contingencies for each link whose product will be identical to the contingency of the overall direct relation linking the initial and final variable of the chain.

However, psychologically it may make a difference whether outside variables are explicitly mentioned or not, or whether interpolated variables are explicitly presented, which possibly makes subjects more likely to consider possible disruptions of these named variables. Research, for example about support theory, has shown that people weigh explicitly mentioned variables more than implicit ones (Tversky & Koehler, 1994). Thus, our hypothesis is that one reason for weakening with indirect relations may be that they invite inferences about possible disruptions triggered by the sequence of variables along the chain.

In the present study we test one aspect of this hypothesis directly. Focusing on disablers, we compare causal models in which disablers are explicitly mentioned with causal models in which they remain implicit. We expect the impact of disablers to be higher when they are explicitly mentioned than when they are only implied. Moreover, we contrast again a direct relation condition with an indirect one with interpolated variables. Our hypothesis predicts a main effect there too because direct relations invite inferences regarding fewer disablers than indirect causal relations which mention more, possibly disruptable variables.

A further goal of the study was to test whether learners solely reason on the basis of causal model information or whether they are actually sensitive to the learning data. As mentioned above, Bes et al. (2012) claimed that subjects largely disregard data. However, because of violations of the Markov condition their learning data were not in line with the instructed causal model, which might have been a reason for neglecting the data. Therefore, we revisited this question by manipulating the size of the contingency presented to learners. Otherwise, we used the same “updating” paradigm as in Experiments 1 and 2. As Experiment 2 did not show any influence of the level of description, we presented only the physiological variables which we already used in Experiment 1.

4.1 Method

4.1.1. Participants

One hundred and ninety subjects (146 female, 41 male, three subjects indicated neither male nor female, $M_{Age} = 34.41$, $SD_{Age} = 11.85$) were recruited via the online platform *Prolific* (www.prolific.ac). This sample size allows it to detect medium main effects and interactions (although we did not predict any interaction effects) of $d = 0.5$ with more than eighty percent probability. All subjects were native English speakers and had at least an A-level degree. They were paid 0.70 £ for their participation.

Table 4

Contingencies presented to subjects in Experiment 3

Contingency	$n(J, L)$	$n(J, \neg L)$	$n(\neg J, L)$	$n(\neg J, \neg L)$	$P(L J)$	$P(L \neg J)$	ΔP
high	18	6	5	19	0.75	0.21	0.54
low	11	13	5	19	0.46	0.21	0.25

4.1.2. Design, Materials, and Procedure

The design was a 2 (causal model: direct vs. indirect) \times 2 (information about disablers: given vs. not given) \times 2 (contingency: high vs. low; see Table 4) between-subjects design. Like in Experiments 1 and 2, subjects first were presented with trial-by-trial learning data presenting a contingency between JPH3 and the disease Lipogastrosis. The contingencies shown to participants are depicted in Table 4. The contingency in the “high contingency” condition was the same as in Experiment 1. The “low contingency” had a lower predictive probability but the same base rate of the effect in the absence of the cause.

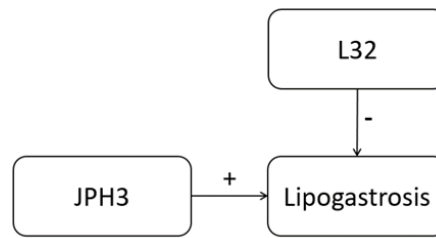


Figure 14. Illustration of the causal model shown to subjects in Experiment 3 in the *direct causal relationship condition* in which disablers were explicitly presented.

After the learning phase, subjects were provided with the causal model information.

The “direct causal relationship” condition in which no information about disablers was given was comparable to the one in Experiment 1. In the condition in which we explicitly mentioned disablers, participants were shown the illustration depicted in Figure 14 and read the following text:

*“Please read the following new information:
 Later, the biologists found out that JPH3 and Lipogastrosis are directly causally related, which is illustrated in the figure below by the arrow that goes from JPH3 to Lipogastrosis. The plus sign above the arrow indicates that the probability of contracting Lipogastrosis is higher for individuals who suffer from a JPH3 mutation compared to individuals who do not have the mutation. In addition, the biologists found out that there also exists a particular gene, L32, that has a protective influence. Specifically, having the gene L32 reduces the probability of Lipogastrosis.”*

Other than in Experiments 1 in which the causal chains contained only one interpolated variable, subjects in the *indirect causal relationship* condition of Experiment 3 were told that the causal model was a four-variable causal chain, with the interpolated variables being Hepatocytosis and Cholestocytis. In the condition in which we explicitly

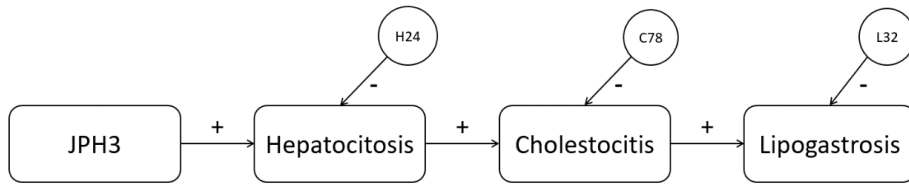


Figure 15. Illustration of the causal model shown to subjects in Experiment 3 in the indirect causal relationship condition in which disablers were explicitly presented.

mentioned disablers, participants were shown the illustration depicted in Figure 15 and read the following text:

*“Please read the following new information:
 Later, the biologists found out that JPH3 and Lipogastrosis are indirectly causally related by a chain that is illustrated in the figure below. Specifically, they found out that individuals who suffer from a JPH3 mutation have a higher probability of contracting Hepatocytosis (indicated by the arrow that goes from JPH3 to Hepatocytosis and the plus sign above the arrow), which is an abnormal increase of hepatic enzymes. Further, for individuals who suffer from Hepatocytosis there is a higher probability of contracting Cholestocytis, an abnormal increase in cholesterol levels. Finally, Cholestocytis increases the probability of contracting Lipogastrosis. In addition, the biologists found out that there also exist different genes, H24, C78, and L32, that have a protective influence. Specifically, having the gene H24 reduces the probability of Lipogastrosis by reducing the probability of Hepatocytosis (indicated by the arrow with the minus sign). Having the gene C78 also reduces the probability of Lipogastrosis by reducing the probability of Cholestocytis. Finally having the gene L32 reduces the probability of Lipogastrosis.”*

After participants were presented the causal model information, they proceeded to the test question. We asked the same predictive probability question as in Experiment 1.

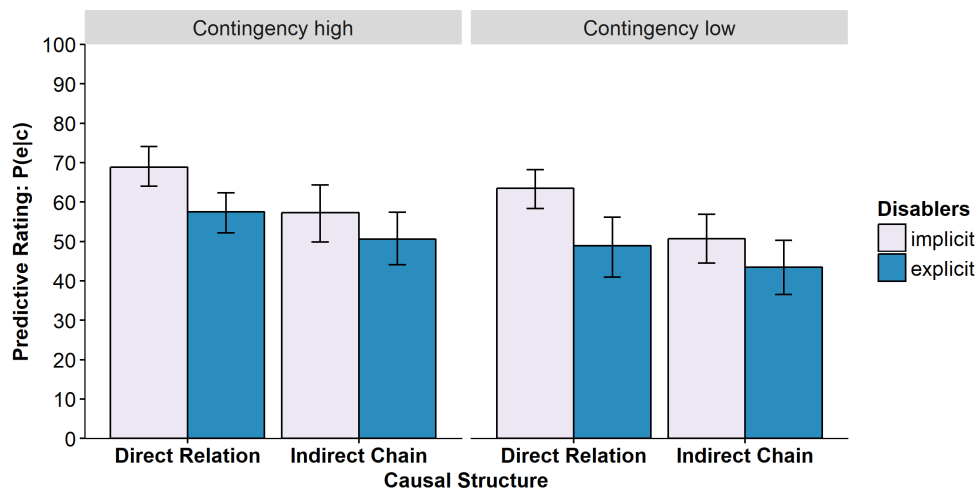


Figure 16. Results (means and 95% CIs) of Experiment 3.

4.2. Results and Discussion

The results are summarized in Figure 16. As can be seen, we replicated again the effect that ratings for indirect causal chains are lower than for direct causal relationships. A $2 \times 2 \times 2$ ANOVA with the between-subjects factors causal model (direct vs. indirect), information about disablers (given vs. not given), and contingency (high vs. low) confirmed that the effect of causal model was significant, $F(1, 282) = 15.80, p < .001, d = .46$. Figure 16 also shows that ratings differed depending on whether subjects were informed about the existence of potential disablers or not. The predictive probability ratings were lower in the condition that explicitly mentioned the existence of disablers than in the condition that did not mention the existence of disablers, $F(1, 282) = 18.69, p < .001, d = .51$. This finding indicates that subjects who are more likely to think about the possibility that disabler may disrupt a causal process tend to lower their predictive probability ratings despite identical covariation data.

We also observed an influence of the contingency that subjects observed. Ratings were overall higher in the condition with the higher contingency, $F(1, 282) = 9.03, p < .01, d$

= .35, which shows that subjects are sensitive to the presented data. The ANOVA yielded no significant interaction effects.

The finding that learners pay attention to the data, contradicts the claim of Bes et al. (2012) that subjects generally ignore learning data. However, we also observed substantial deviations from the presented probabilities, some of which might be explained by regression effects (see Rottman & Hastie, 2014, for evidence for regression to the mean in causal model learning). Thus, subjects clearly weigh causal model information more strongly than the observed data. Under the assumption that causal model information probably relies on stronger evidence than a single study, weighing causal model instruction more strongly seems reasonable.

The experiment also shows that explicit information about disablers leads to weaker predictive probability ratings than implicit information, despite identical contingencies. Thus, explicitly mentioning disablers leads to a further weakening of predictive probability estimations beyond what we have seen so far. We focused on disablers but would have expected a similar finding had we pointed out the possibility of an absence of enablers.

We also found an independent weakening effect. Again, direct relations yielded higher predictive probability ratings than indirect relations. This finding is also consistent with our hypothesis if it is assumed that the likelihood of thinking about disablers increases with the number of variables to be considered in both the explicit and the implicit conditions. However, we do not have direct evidence that this is in fact underlying the weakening effect in the two conditions, the obtained empirical pattern is only consistent with the hypothesis.

The hypothesis that multiple variables on a chain should lead to weakening because more disablers might be considered in the indirect representation is actually only valid under the assumption that subjects do not represent causal strengths along the links correctly in

interpolation scenarios. Since in interpolation scenarios the product of the strengths of the links of the interpolated chain should be identical to the strength of the overall contingency between the two variables linked in both the direct and indirect causal relationship condition, both representations should reflect the same impact of disablers. If subjects normatively represented interpolated chains, the impact of disablers on the direct causal relation will be identical to the sum of impacts of the disablers along the chain links. However, to accomplish this, subjects need to understand that the causal strengths of individual links become stronger, the more links are interpolated (see section 5 for an explanation). If subjects do not understand this property, and do not modify link strengths relative to the length of the chain, then additional interpolated links will indeed increase the represented impact of disablers and hence lead to a weakening effect. The next experiment will directly test subjects' beliefs about the strengths of links in interpolation scenarios.

5. Experiment 4

The goal of Experiment 4 was to test more directly subjects' belief of the causal strengths of interpolated links. The key idea is that people may have difficulties differentiating between lengthening and interpolations. On the surface, the results of both processes look the same. One cannot tell just by looking at a chain with several variables whether it is the result of a lengthening or an interpolation process. Moreover, lengthening may be the more frequent process in daily life. For example, when planning interventions, we may focus on short-term or long-term effects. Long-term effects are less likely than short-term effects. By contrast, interpolations occur most frequently in the context of discovery and learning. It may well be that this is a less frequent scenario in daily life. Given that lengthening often leads to a weakening effect, confusing lengthening with

interpolations and focusing on lengthening explains why we see an erroneous weakening effect also in interpolations.

One way to test this hypothesis is by focusing on the probabilistic relations people assume for the individual links within a causal chain. In lengthening scenarios, the process starts with a causal chain (or a direct causal relation), which stays fixed. Additional variables are added at either side of the chain with flexible strength. According to the multiplication rule (Equation 1) this should in probabilistic chains typically lead to a weakening effect. Importantly, the strengths of the initially represented causal links are not affected by the addition of further links extending the original model.

By contrast, in an interpolation scenario, the covariation between the initially presented cause C and effect E should be more or less stable (depending on the reliability of the measurements). When interpolating variables, the strengths of the newly introduced links should go up proportional to the number of interpolated links. This is a basic consequence of the multiplication rule. Given that a multiplication of causal strength estimates for the links of the chain need to result in an invariant value for the initially presented C and E, their values need to go up to result in an equal result of the product.

If, however, people confuse interpolations with lengthening, we expect to see relatively invariant strength estimates for the individual links regardless of the length of the interpolated chain. The size of the link estimates may be influenced by the initially observed covariation and/or some strength prior. If invariance of link strength estimates is assumed, then, due to the multiplication rule, a weakening effect should occur here as well.

To test our theory, we again used our standard interpolation task, while contrasting different conditions in which the number of interpolated variables was manipulated. The contingency for the initially presented cause and effect is shown in Table . In this

experiment, we used a contingency which implies that JPH3 is a necessary cause of Lipogastrosis (i.e., $P(L|\neg J)=0$). This simplifies the calculation because $P(L|J)$ should simply be the product of the predictive probabilities of its components. However, given that the multiplication rule (Equation 1) refers to contingencies (ΔP) and we know from previous research that subjects often provide a more conservative positive estimate when asked about the probability of an effect in the absence of the cause, $P(L|\neg J)$, we also had subjects estimate this quantity. We did not expect that this estimate varies with conditions so that our general prediction should be unaffected.

We contrasted a *direct causal relationship* condition with two *indirect causal relationship* conditions (a two-links and a four-links chain). As dependent variables we asked subjects to give an average estimate of the probabilistic relations of the individual links. Importantly, no data were presented about the individual interpolated links. As in most of our studies, the causal chains were just verbally instructed. Normatively, the averaged estimates should go up with increased length of the chains (see the line with the normative predictions in Figure 19 and Table 5). If, by contrast, learners confuse lengthening with interpolations, we expect relatively invariant estimates.

Table 5

Contingency presented to subjects in Experiment 5

	L	¬ L	$P(L J)$	$P(L \neg J)$	ΔP	$M \Delta P_{n=2}$	$M \Delta P_{n=4}$
J	10	14	.42	.0	.42	.65	.81
¬ J	0	24					

Note $M \Delta P_{n=2}$ and $M \Delta P_{n=4}$ denote the average single-link contingencies for a causal chain with a distal contingency of ΔP and $n = 2$ vs. $n = 4$ known links.

5.1 Method

5.1.1. Participants

Two hundred and ten subjects (124 female, 82 male, four subjects indicated neither male nor female, $M_{Age} = 35.55$, $SD_{Age} = 11.29$) were recruited via the online platform *Prolific* (www.prolific.ac). This sample size allows it to detect even a small interaction effect of $d = 0.25$ between test query and causal structure. A strong interaction effect is predicted by the normative values but we hypothesized that we would see no effect. To have a strict test, we tested therefore against the possibility of a small interaction effect. All subjects were native English speakers and had at least an A-level degree. They were paid 0.70 £ for their participation.

5.1.2. Design, Materials, and Procedure

The design was a 3 (causal model: direct relationship vs. two-links chain vs. four-links chain; between subjects) \times 2 (type of probability query: $P[\text{variable} | \text{parent}]$ vs. $P[\text{variable} | \neg\text{parent}]$; within-subject) mixed design. The overall procedure was comparable to the ones used in the previous experiments. However, since the task is slightly more complex, we tried to increase attentional involvement during learning by changing the initial learning task into an active supervised learning paradigm. An illustration of the task is shown in Figure 17. We first showed subjects only the cause status for every of the total 48 observations, and then asked them to make a guess about the status of the effect. To check whether the effect was present or absent, subjects had to click on a “check disease” button

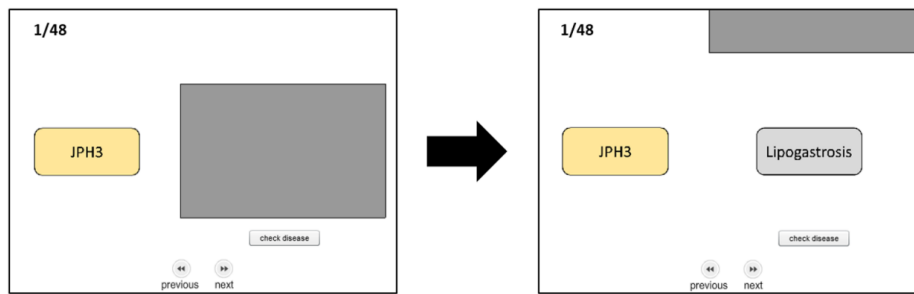


Figure 17. Illustration of the learning task used in Experiment 4. The left picture shows the initial state of the screen, at which only the status of the cause factor was shown. The effect status was only revealed after subjects clicked on the “check disease” button displayed below a grey mask. The right picture shows what the screen looked like when the effect status was revealed to participants. The illustration depicts a case in which the cause was present (yellow JPH3 box) but the effect was absent (gray Lipogastrosis box). With the “previous” and “next” buttons, subjects were able to navigate through the sample (consisting of 48 cases in total).

displayed on the screen (see Figure 17). Another difference from the previously used paradigm was that subjects had to navigate actively through the observations in the sample. They had to click on a “next” button to proceed to subsequent cases and could also click on a “previous” button if they wanted to go back to earlier cases. For each participant, the order in which the 48 cases were shown was randomly determined prior to the presentation of the first case of the sample.

As in most previous experiments, the causal model information was given to participants after they had finished the learning task. The information shown to participants in the *direct causal relationship* and in the *two-links chain* condition was like the one we used in previous experiments. In the *four-links chain* condition, subjects were presented with the following text and the illustration shown in **Fehler! Verweisquelle konnte nicht gefunden werden**.¹⁸:



Figure 18. Illustration of the causal model shown to subjects in Experiment 4 in the four-links chain condition. The order of the middle variables Hepatocytosis, Spirillicytis, and Cholestocytis was randomized between subjects.

“Please read the following paragraph which provides a new piece of information: The biologists later found out that JPH3 and Lipogastrosis are in fact indirectly causally related as illustrated in the figure below. Specifically, the JPH3 mutation can sometimes lead to Hepatocytosis, an abnormal occurrence of hepatic enzymes. This is indicated by the arrow (with a + sign) that goes from JPH3 to Hepatocytosis. Furthermore, Hepatocytosis can sometimes lead to Spirillicytis, an infection with a bacterium colonizing the gut. This is indicated by the arrow (with a + sign) that goes from Hepatocytosis to Spirillicytis. Furthermore, Spirillicytis can sometimes lead to Cholestocytis, an abnormal occurrence of cholesterol. This is indicated by the arrow (with a + sign) that goes from Spirillicytis to Cholestocytis. Finally, Cholestocytis can sometimes lead to Lipogastrosis. This is indicated by the arrow (with a + sign) that goes from Cholestocytis to Lipogastrosis. Other factors can also influence the disease.”

Other than in the previous experiments, the order of the interpolated variables in the *four-links chain* (Hepatocytosis, Spirillicytis, and Cholestocytis, see Fehler! Verweisquelle konnte nicht gefunden werden.18) condition was randomized between subjects. The labels of the interpolated variable in the *two-links chain* condition was randomly varied between subjects and could be either of the interpolated variable labels we used in the *four-links chain* condition.

After subjects were instructed about the underlying causal model, they proceeded to the inference screen on which we asked them to make two ratings assessing the assumed strength of a single link of the introduced causal model. Given that no new data was presented so that subjects could not possibly distinguish between links, we instructed them that they should assume that all links are equally strong, and were only asked about one randomly picked link.

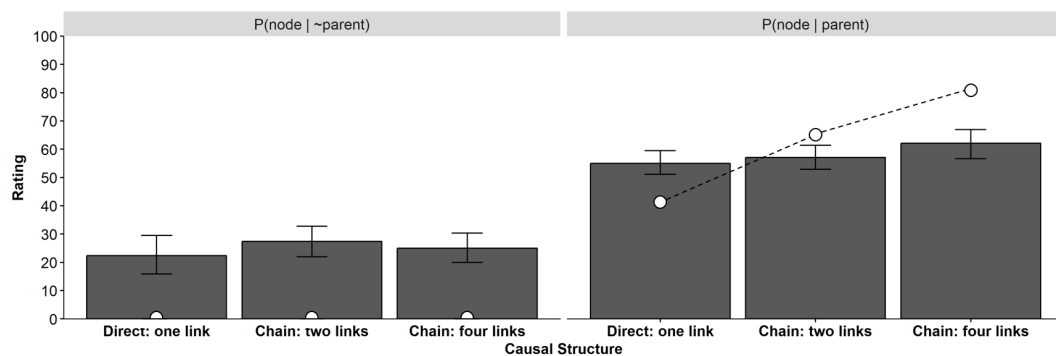


Figure 19. Results (means and 95% CIs) of Experiment 4 and normative values shown in the dotted line.

For example, in the *four-links chain* condition subjects first were presented with the following text, accompanied again by the illustration of the causal model:

“The following questions refer to the causal chain between JPH3 and Lipogastrostis. Please assume that the four causal links (expressed by the four arrows) in the graph represent equally strong causal relations.”

Subsequently, we asked subjects to estimate the probability that a particular effect of the introduced causal model was present given that its direct parent variable was present, $P(\text{variable} | \text{parent})$, as well as the probability that this effect was present if the direct parent variable was absent, $P(\text{variable} | \sim\text{parent})$. For example, some participants in the *four-links chain* condition were asked the following two questions:

“Assuming that all links are equally strong, how likely do you think is it, for example, that a mouse having Hepatocytosis has Spirillicitis?” And *“Assuming that all links are equally strong, how likely do you think is it, for example, that a mouse not having Hepatocytosis has Spirillicitis?”* Ratings for both questions were provided on a slider ranging from 0 to 100 (with endpoints labelled *“It is certain that this mouse does not have Spirillicitis.”* and *“It is certain that this mouse has Spirillicitis.”*). The specific link to which the test questions referred in the *two-links chain* and the *four-links chain* condition was randomized between subjects.

5.2. Results and Discussion

The results are summarized in Figure 19 (means and 95% CIs). The left part of Figure 19 shows the results for the base rate ratings in the absence of the cause, $P(\text{variable} | \neg\text{parent})$, which in the *two-links* and *four-links chain* conditions were averaged over the respective link of the causal model subjects were asked about. The right part shows the results for the predictive probability estimations, $P(\text{variable} | \text{parent})$. The dotted line depicts the normative (average) values for these parameters. As can be seen, subjects gave overall higher ratings when they were asked to estimate $P(\text{variable} | \text{parent})$ than when they were asked to estimate $P(\text{variable} | \neg\text{parent})$, which corresponds to what they have seen. However, it can also be seen that the $P(\text{variable} | \neg\text{parent})$ ratings were notably higher than the normative values (which were 0). However, they stayed relatively invariant across the conditions, which allows us to focus on the predictive probabilities ($P(\text{variable} | \text{parent})$), subjects were asked about.

Figure 19 shows that averaged ratings for the $P(\text{variable} | \text{parent})$ estimations descriptively indicate no difference between the direct and indirect two-links chain conditions. There is a hint of an upward trend for the four-links conditions but the estimates are much closer to the other estimates than the normative value (see Figure 19). A 3 (causal model: direct relationship vs. two-links chain vs. four-links chain; between subjects) \times 2 (type of probability query: $P[\text{variable} | \text{parent}]$ vs. $P[\text{variable} | \neg\text{parent}]$; within subject) mixed ANOVA only yielded a significant main effect of “type of probability query”, $F(1, 207) = 224.73, p < .001, d = 1.47$, confirming that $P(\text{variable} | \text{parent})$ ratings were indeed higher than $P(\text{variable} | \neg\text{parent})$ ratings. Importantly, there was no significant interaction between “causal model” and “type of probability query,” which would have been expected based on

the normative values. Post-hoc tests (Scheffé test) revealed that the $P(\text{variable} | \neg\text{parent})$ ratings as well as the $P(\text{variable} | \text{parent})$ ratings did not differ from each other.

In sum, the results of this experiment are consistent with our prediction that subjects have a hard time understanding the normative implications of causal interpolations, and may not clearly understand the difference between lengthening and interpolations. This is understandable because the causal models look the same in the two cases and do not reveal the process underlying their construction. Consequently, subjects are largely insensitive to the normative implication of increased strength with an increased number of interpolated links. Overall, the results are consistent with the theory that subjects tend to assume fairly stable link strengths, as in lengthening scenarios. Keeping the strengths of the links relatively invariant, entails a weakening effect proportional to the length of the chain, which is what we have seen in the previous studies. The results of the experiment are consistent with the theory presented in the context of Experiment 3 that increased length of the interpolated chain increases the potential impact of disablers both when the disablers are implicitly (in the causal strength parameters) and explicitly represented.

6. Experiment 5

In the previous experiments we focused on the contrast between direct causal relations and indirect causal chains with various lengths. Chains are only one example for a mechanism linking two variables. The two variables may also be mediated by more complex causal models, for example, by causal diamond structures which link cause and effect through two parallel converging chains. Experiment 5 focuses on such more complex causal models.

Normatively, there is again an asymmetry between a situation in which a cause-effect relation is later augmented by embedding it into a complex causal model (analogous to

lengthening), and an interpolation of a complex causal model as a representation of a newly discovered mechanism linking two previously observed variables. For the same reasons as with simple chains, interpolations of a complex causal models should not systematically alter the covariation between the two variables. Again, the most plausible assumption is that the initially observed covariation between the two variables had already been mediated by a complex causal model, fragments of which were only later discovered.

By contrast, if the same causal model (e.g., diamond structure) is generated as a result of an augmentation process (the initially presented two variables would have to differ, of course), then the predictive probabilities and contingencies between the events would be dependent on the structure of the causal model and its parameter (i.e., causal strengths, functional form).

Consider the causal models in Figure 20, which we have used in the present experiment. If the covariation between JPH3 and Lipogastrosis has been reliably established in the initial learning phase, interpolating any of the four causal models should not systematically alter this covariation (beyond what is expected due to sampling variation).

Now, for examples of augmentations, we can use the causal models in Figure 20b, c, or d. Assume, for example, that the process starts with the JPH3-H24 relation. If later the other variables are being discovered and added to the initial causal model fragment, the probabilistic relationship between JPH3 and Lipogastrosis will crucially depend on the structure of the causal model (Fig. 20b, c, or d), the causal strengths of the added links, and the functional form. As for functional form, given that the graph conveys that the common cause model and the two chains emanating from JPH3 in Figure 20c and d honors the Markov condition, the open question here is how the common effect model converging on the effect Lipogastrosis combines the two causes, PABA and TAMP. We compared a standard

disjunctive model (noisy-OR) (Fig. 20c) in which each cause individually and both causes together probabilistically cause the effect with a conjunctive model (noisy-AND)(Fig. 20d) in which both causes have to be present to probabilistically cause the effect.

What do we predict in the present experiment? Given that Experiment 4 suggests that learners tend to treat interpolations like augmentations (e.g., lengthening), we expect that the ratings for the JPH3-Lipogastrosis relation will be affected by the mediating causal model. We should again see a weakening effect because the two events are mediated by two chains. However, depending on whether the two chains are viewed as alternative routes to the effect, thus compensating each other for failures of causation (disjunctive noisy-OR) or as a conjunctive Noisy-AND structure which requires both causes to be present, we predict different sizes of the weakening effect. The lowest ratings should be seen in the conjunctive Noisy-AND model because the effect seems less likely when both causes need to be present.

6.1 Method

6.1.1. Participants

Three hundred and twenty-nine subjects (208 female, 121 male, $M_{Age} = 35.01$, $SD_{Age} = 10.77$) were recruited via the online platform *Prolific* (www.prolific.ac). This sample size allows it to detect a small main effect of $d = 0.40$ for the causal model factor. All subjects were native English speakers and had at least an A-level degree. They were paid 0.70 £ for their participation.

6.1.2. Design, Materials, and Procedure

One factor with three levels (type of causal model: *direct link vs. causal chain vs. conjunctive diamond vs. disjunctive diamond*) was manipulated between subjects. Subjects were randomly allocated to the different conditions. We used the same learning paradigm as

we did in Experiments 1, 2, and 3. Thus again, in Phase 1 subjects learned about a covariation between JPH3 and Lipogastrosis. We used the same learning data as in Experiment 1 (see Table 1). The probability of Lipogastrosis given a mutation of JPH3 was therefore again $P(L|J) = 0.75$ and the base rate of Lipogastrosis in the absence of Lipogastrosis was $P(L|\neg J) = 0.21$. Hence, the contingency was $\Delta P = 0.54$.

Next, the causal model instructions were presented (see Table 6 for the different conditions). The graphic models that were presented together with the instructions are shown in Figure 20.

Table 6

Causal-model instructions presented in the different conditions of Experiment 5.

Condition	Instruction
Direct link	After having completed their study, the biologists hypothesized that JPH3 and Lipogastrosis are directly causally related, as illustrated in the figure below. Additional data confirmed the biologists' hypothesis. Specifically, mice who carry the JPH3 mutation have an increased probability of contracting Lipogastrosis. This is indicated by the arrow (with a + sign) that goes from JPH3 to Lipogastrosis.
Causal chain	After having completed their study, the biologists hypothesized that JPH3 and Lipogastrosis are indirectly causally related as illustrated in the figure below. Additional data confirmed the biologists' hypothesis. Specifically, mice who carry the JPH3 mutation have an increased probability of expressing a particular other gene, H24 [C78]. This is indicated by the arrow (with a + sign) that goes from JPH3 to H24 [C78]. Further, mice who express H24 [C78] have a higher probability of producing the enzyme PABA [TAMP]. This is indicated by the arrow (with a + sign) that goes from H24 [C78] to PABA [TAMP]. Finally, mice who produce the enzyme PABA [TAMP] have a higher probability of contracting Lipogastrosis. This is indicated by the arrow (with a + sign) that goes from PABA [TAMP] to Lipogastrosis.
Disjunctive diamond	After having completed their study, the biologists hypothesized that JPH3 and Lipogastrosis are indirectly causally related, as illustrated in the figure below. Additional data confirmed the biologists' hypothesis. Specifically, mice who carry the JPH3 mutation have a higher probability of expressing two particular other genes, H24 and C78. This is indicated by the arrows (with a + sign) that go from JPH3 to H24 and C78. Further, mice who express H24 have a higher probability of producing the enzyme PABA. Likewise, mice who express the gene C78 have a higher probability of producing the enzyme TAMP. This is indicated by the arrows (with a + sign) that go from H24 to PABA and from and C78 to TAMP. Finally, mice who produce the enzymes PABA and [in conjunctive]/or [disjunctive] TAMP have a higher probability of contracting Lipogastrosis. This is indicated by the arrows (with a + sign) that go from PABA and TAMP to Lipogastrosis. Importantly, both PABA and TAMP independently of each other increase the probability of contracting Lipogastrosis. Hence, mice with either PABA or TAMP have an increased probability of contracting the disease, and mice who have both PABA and TAMP have the highest probability of contracting the disease.
Conjunctive diamond	After having completed their study, the biologists hypothesized that JPH3 and Lipogastrosis are indirectly causally related, as illustrated in the figure below. Additional data confirmed the biologists' hypothesis. Specifically, mice who carry the JPH3 mutation have a higher probability of expressing two particular other genes, H24 and C78. This is indicated by the arrows (with a + sign) that go from JPH3 to H24 and C78. Further, mice who express H24 have a higher probability of producing the enzyme PABA. Likewise, mice who express the gene C78 have a higher probability of producing the enzyme TAMP. This is indicated by the arrows (with a + sign) that go from H24 to PABA and from and C78 to TAMP. Finally, mice who produce the enzymes PABA and [in conjunctive]/or [disjunctive] TAMP have a higher probability of contracting Lipogastrosis. This is indicated by the arrows (with a + sign) that go from PABA and TAMP to Lipogastrosis. Importantly, the probability of contracting Lipogastrosis is only increased when both PABA and TAMP are present. Hence, mice who have only PABA or TAMP do not have an increased probability of contracting the disease. Only mice who have both PABA and TAMP have an increased probability of contracting the disease.

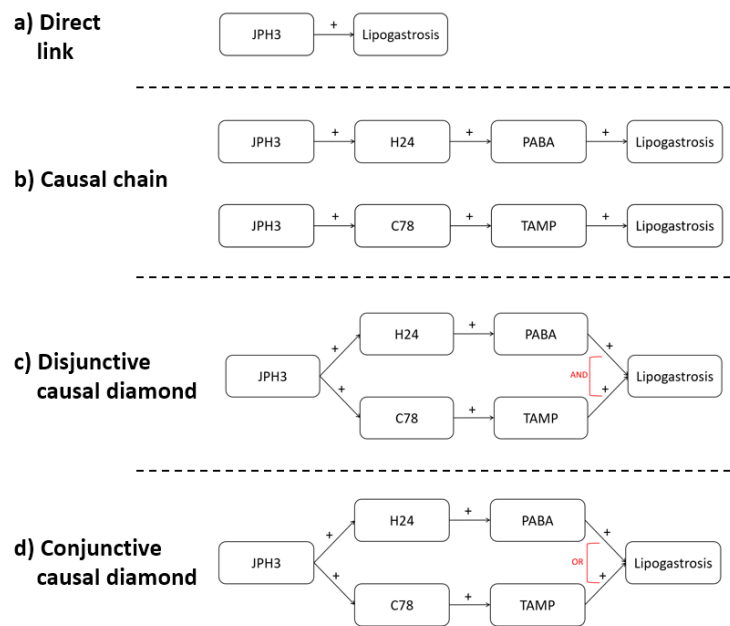


Figure 20. Causal graphs shown in the different conditions of Experiment 5. Whether subjects were presented the upper or lower causal chain in the causal-chain condition was randomized. We did this because we wanted to present subjects with the same labels and variable descriptions that were used in the two causal-diamond conditions.

Figure 20b shows that we used two different sets of interpolated variables in the causal chain condition. Whether the interpolated variables were “H24” and “PABA” or “C78” and “TAMP” was randomly altered between subjects. We used these two different sets of interpolated variables because they corresponded to the variables that were described in the two causal diamond conditions.

After subjects had studied the causal model information, they proceeded to the test question. Like in previous experiments, subjects read that the biologists had randomly sampled a new mouse that carries the JPH3 mutation. We also showed the respective causal model again, with the JPH3 variable being marked in yellow. Subjects then were asked to estimate the probability that this mouse would also be suffering from Lipogastrosis. Ratings

were provided on a slider from 0 to 100, with the endpoints labeled as “it is certain that this mouse does not have Lipogastrosis” and “it is certain that this mouse has Lipogastrosis.”

6.2. Results and Discussion

The results are summarized in Table 7. Ratings in the *causal chain* condition were again lower than the ratings in the *direct link* condition, replicating the weakening effect. However, the difference was smaller than in previous experiments. The ratings in the conjunctive and disjunctive diamond conditions were also lower than the ratings in the direct link condition, implying that the weakening effect does occur for other indirect causal models, too. As predicted, ratings also differed between the conjunctive and the disjunctive diamond conditions. When subjects had learned that two factors needed to be present for the effect (conjunctive case), their predictive probability estimations tended to be lower than when they had learned that either of the two different factors independently can cause the effect (disjunctive case).

Table 7
Summary of the results of Experiment 5.

	Direct	Causal chain	Disjunctive diamond	Conjunctive diamond
Mean	66.30	60.83	58.82	53.00
<i>SD</i>	18.31	19.95	18.58	19.54
Median	70.00	60.00	60.00	52.00
95% <i>CI</i>	4.05	4.33	4.06	4.32

The statistical analyses partially confirm these impressions. First of all, a global one-way ANOVA was significant, $F(3,325) = 6.69$, $p < .001$, $d = .51$, confirming that subjects did indeed make different predictive probability judgments in the four different causal model conditions. Planned contrasts (two sided) revealed that the difference between the *direct link* and the *chain* condition was not significant this time, $t(325) = 1.84$, $p = .07$. The weakening effect was smaller than in our previous studies, $d = 0.29$. If we compare the

means shown in Table 7 with the ones obtained in Experiment 1 we see that the smaller difference between the direct and the causal-chain condition was not solely due to higher ratings in the causal-chain condition. The ratings in the direct link condition were also lower than those in Experiment 1.

The difference between the *direct link* and the *disjunctive diamond* condition was also significant, $t(325) = 2.51, p = .01, d = 0.41$. The difference between the *conjunctive* and the *disjunctive diamond* conditions missed significance, $t(325) = 1.95, p = .05, d = 0.31$. Ratings in the *conjunctive diamond* condition were significantly lower than the ratings in the *direct link* condition, $t(325) = 4.43, p < .001, d = 0.70$. Finally, ratings in the *conjunctive diamond* condition were lower than ratings in the *causal-chain* condition, $t(325) = 2.63, p < .01, d = 0.40$. We used two-sided tests as in the previous experiments but one could defend one-sided tests for the tests of the conjunctive diamond condition here, which yielded the lowest ratings as predicted.

In sum, this exploratory experiment shows again that interpolations lead to weakening effects, which now could also be found with more complex interpolated causal models. Normatively, no difference whatsoever should have been observed between causal model conditions, had subjects a full grasp of interpolations. Despite some weak (or possibly non-existent) effects in the individual comparisons, the strong main effect of the causal model factor clearly supports the hypothesis that people do not fully understand interpolations and rather treat them like augmentations. Further specific support for this hypothesis is provided by the results of the conjunctive diamond model condition. This model seemed to convey that the effect should be less likely which is then reflected in somewhat lowered ratings. Of course, this study requires follow-up work but it is clearly consistent with what we have seen with causal chains.

7. General Discussion

Causal knowledge is not static, it is constantly modified and changed based on new evidence. The present set of studies explores one important case of causal belief revisions, causal interpolations. The prototypic case of an interpolation is a situation in which we initially have gathered knowledge about a direct causal or covariational relation between two variables but later become interested in the mechanism linking these two variables. We may, for example, use our knowledge about one effect of a drug such as Aspirin in daily life but later are interested in learning how Aspirin exerts its effect. Interpolations in science are frequent. Scientists have, for example, in the past decades traced the precise mechanisms initiated by the intake of Aspirin. This knowledge has led to knowledge about further effects of the drug beyond relieving headaches (e.g., prevention of heart attacks in specific populations).

Our key finding of our research is that interpolations tend to be misrepresented, which leads to the paradox of knowing more: The more we know about the mechanism, the less predictable we seem to find the effect we are trying to predict. Interpolating a mechanism should normatively not alter the statistical relationship between the variables that initially are represented as direct. This can be best understood by considering what role the mechanism plays prior to its discovery. The discovery of a mechanism linking two variables does not change the causal relation. If the discovery correctly identifies variables on the mechanism path, these variables were already there and effective prior to their discovery (see Fig. 4). Interpolations do not add mechanisms, they just discover them. Thus, if the covariation between two variables has been mediated by the later discovered mechanism all along, discovering fragments of the mechanism should not alter the covariation, unless new data has been collected that contradict the initial findings.

By contrast, we found a systematic *weakening* effect when a mechanism (e.g., a causal chain or a more complex causal model) was interpolated. The predictive probability tends to be estimated significantly lower for two variables when a mechanism is interpolated than when they are represented as a direct relation. We replicated this effect in various conditions with different types of variables and different types of interpolated causal structures. We also found that the effect is so intuitive that even experts (which includes co-authors) were initially convinced that weakening is normative.

To provide a psychological explanation for the weakening effect, we pursued two interrelated hypotheses, which both were supported. The overarching theoretical explanation behind both hypotheses is that people have difficulties distinguishing between interpolations and augmentations. If given chains, for example, are lengthened at the outer ends of the chain, then a weakening of the predictive relation between the initial cause and the added probabilistic effects is in most cases normatively expected. The longer the probabilistic chain, the weaker the statistical relations tend to become. By contrast, in interpolations the covariation between the two initially directly linked variables should stay invariant regardless of the length of the interpolated chain. However, our evidence shows that people seem to confuse interpolations with lengthening and therefore activate intuitions about lengthening to make predictions in interpolation scenarios. This is understandable if it is assumed that lengthening is more frequent in daily life. Moreover, on the surface a chain that is generated by lengthening cannot be distinguished from one that is the result of interpolations. Apparently, subjects are mainly sensitive to the ordinal length of chains regardless of how they were generated.

Confusing interpolations with lengthening (or other types of augmentations of causal knowledge) entails specific interrelated predictions, two of which we have tested. The first

claims that the greater number of variables in a chain should sensitize subjects to a larger number of possible disruptions (i.e., disablers, lack of enablers, alternative causes) whereas in direct relations all possible disruptions are condensed in the representation of a single link (and a summarized alternative cause). Normatively, in cases of interpolation it should actually not make a difference whether a covariation is represented as a consequence of direct causation or as indirect. The direct relationship should normatively lump together all the possible disruptions of the interpolated chain. However, this property requires an understanding of the fact that the causal strengths of the links of the chains need to be adapted to the length of the chain. If, by contrast, interpolation is confused with lengthening, then it is indeed true that disruptions (and hence weakening) become more potent the longer the chain is. We predicted that subjects' predictive probability estimates about the initially presented two covarying events will be affected by the greater length of the interpolated chain because we predicted that subjects tend to represent more potentially disrupting variables for chains compared to direct causal relations. Supporting this hypothesis, we generally found that direct causal relations received higher predictive probability ratings than indirect causal relations. As second evidence for the disruption hypothesis we have shown in Experiment 3 that weakening becomes even stronger when disablers are explicitly labelled. This is consistent with research showing that explicit representations are outweighed relative to implicit ones (Tversky & Koehler, 1994).

The hypothesis that longer chains highlight more possible disruptions implies that subjects do not understand that in interpolations the causal strengths of links should be estimated as stronger, the more mechanism components are added. Given that the product of contingencies of the interpolated links should equate the overall contingency between the two variables in the direct representation, increasing the resolution of the interpolated

chain should lead to an average increase of the link contingencies. The more is known about the steps of the mechanism, the more the causal strength estimates for the interpolated links should approximate determinism. If, by contrast, interpolations are misrepresented as lengthening, then it should be expected that the estimates of the causal strengths of the links are independent of the length of the interpolated chain. If causal strength estimates largely stay invariant regardless of the length of the interpolated chain, a weakening effect is predicted as in lengthening scenarios. Our evidence in Experiment 4 supports this hypothesis.

Perspectives for Future Research

Our main focus in the present research was on comparing direct relationships with interpolated chains. Experiment 5 adds to these studies the first examination of more complex causal models. It would be interesting to further study more complex mechanisms and explore moderators of the weakening effect. It would, in particular, be interesting to explore conditions in which the weakening effect is reversed.

We have several times pointed out that our claim that interpolated mechanisms should not alter our probabilistic beliefs about the linked two variables only strictly applies if it is assumed that the context of covariation learning does not differ from the context in which interpolations were made. One advantage of mechanism knowledge is that it provides us with more precise knowledge about what we should expect when we make predictions in new contexts. If, for example, the interpolated chain contains more fine-grained representations of relevant enablers, disablers, and alternative causes, this should allow us to make more precise predictions when we enter a new context. Context changes may lead to an increase of weakening, but strengthening may also be conceivable in both direct and indirect relations. It is important to note that context changes do not generally predict a

weakening effect. If the new context lacks an enabler that was present in the learning context, for example, weakening is predicted, if an enabler is present in the new context that was missing in the learning context, strengthening should be observed. Direct causal relations implicitly represent disablers, enablers, and alternative causes as well, but there these factors are lumped together, which makes the results of transfer harder to predict.

Another interesting question concerns partial knowledge. Imagine we gain knowledge about parts of a complex mechanism but have a hunch that there are other components we do not know yet (e.g., alternative paths to the effect); it would be interesting to see how predictions are altered when subjects become aware that their mechanism knowledge is partial and fragmentary.

Another possibility for future research is to explore the interaction between predictability assessments and prior knowledge. In discussions with colleagues we were sometimes confronted with cases which seem to indicate that mechanism knowledge may make a poorly understood initially direct causal relation more plausible, thus possibly reversing the weakening effect.¹ For example, understanding how a vaccine helps us may make its efficiency more plausible than before. We have focused on fairly abstract materials for which no prior knowledge was available to be able to study the impact of causal structure information that is relatively uncontaminated by prior knowledge. But of course, it is well established that prior knowledge affects our probability assessments. Knowledge about a causal mechanism typically increases estimates of perceived correlations relative to the objective covariation in the data (see, e.g., Fugelsang & Thompson, 2003; Koslowski, 1996; Perales, Shanks, & Lagnado, 2010). Based on these findings, we may find a reversal of the weakening effect if the initial direct causal relation seems implausible according to our

¹ We thank Aaron Blaisdell and Keith Holyoak for mentioning this possibility.

prior knowledge, but the interpolated links raise the plausibility of the causal relation. This mismatch between the direct relations and the knowledge about interpolated links might not be a particularly frequent case, but it is certainly conceivable.

Another more general question is whether weakening effects may tamper our urge to gain more knowledge. Recent research has shown that people tend to know very little about mechanisms (Rozenblit & Keil, 2002; Sloman & Fernbach, 2017). Although it is rational to believe that knowing more increases the predictability and controllability of the world, some people might, based on erroneous intuitions about weakening effects, be reluctant to find out more about mechanisms. Causal knowledge revision is certainly an important topic for causal reasoning research and should attract more interest in the future.

References

- Ahn, W., & Dennis, M. (2000). Induction of causal chain. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 19 – 24). Mahwah, NJ: Erlbaum.
- Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 153.
- Bes, B., Sloman, S., Lucas, C. G., & Raufaste, É. (2012). Non-Bayesian inference: Causal structure trumps correlation. *Cognitive Science*, 36, 1178-1203.
- 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, 301 – 338.
- Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121, 463 – 480.
- Dowe, P. (2007). *Physical causation*. New York: Cambridge University Press.
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140, 168 – 185.
- Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, 31, 800-815.
- Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115, 225 – 237.
- Jara, E., Vila, J., & Maldonado, A. (2006). Second-order conditioning of human causal learning. *Learning and Motivation*, 37, 230 – 246.
- Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, 116, 856 – 874.

- Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, 107, 397 – 402.
- Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: The MIT Press.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955 – 984.
- Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, 39, 65-95.
- Mayrhofer, R., & Waldmann, M. R. (2016). Sufficiency and necessity assumptions in causal structure induction. *Cognitive Science*, 40, 2137-2150.
- Nagel, J., & Waldmann, M. R. (2016). On having very long arms: How the availability of technological means affects moral cognition. *Thinking & Reasoning*, 22, 184-208.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
- Perales, J. C., Shanks, D. R., & Lagnado, D. (2010). Causal representation and behavior: The integration of mechanism and covariation. *Open Psychology Journal*, 3, 174-183.
- Perales, J. C., Catena, A., Cándido, A., & Maldonado, A. (2017). Rules of causal judgment: Mapping statistical information onto causal beliefs. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 29 – 52). New York: Oxford University Press.
- Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: inferences on causal networks. *Psychological Bulletin*, 140, 109 – 139.

- Rottman, B. M. (2017). The acquisition and use of causal structure knowledge. In M.R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 85 – 114). New York: Oxford University Press.
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521-562.
- Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. New York: Riverhead Books.
- Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. Cambridge, MA: The MIT Press.
- Spohn, W. (2012). *The laws of belief: Ranking theory and its philosophical applications*. New York: Oxford University Press.
- von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory & Cognition*, 44, 469-487.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547 – 567.
- von Sydow, M., Hagmayer, Y., Meder, B., & Waldmann, M. R. (2010). How causal reasoning can bias empirical evidence. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2087-2092). Austin, TX: Cognitive Science Society.
- Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 733-752). New York: Oxford University Press.
- Waldmann, M. (Ed.). (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

- Yeung, S., & Griffiths, T. L. (2011). Estimating human priors on causal strength. In L. Carlson, C. Hölscher & T. Shipley (Eds.) *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, 76, 1-29.

Appendix H

Nagel and Stephan (2015)

Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1691 – 1696). Austin, TX: Cognitive Science Society.

Mediators or Alternative Explanations: Transitivity in Human-Mediated Causal Chains

Jonas Nagel (jnagel1@uni-goettingen.de)

Simon Stephan (sstepha1@uni-goettingen.de)

Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

We investigate how learning that an established type-level causal relationship is implemented by human agency affects people's conceptualization of this relationship. In particular, we ask under what conditions subjects continue to perceive the original root cause as appropriate explanation for the resulting effect, and under what conditions they perceive the mediating intentional action as alternative explanation instead. Using a new experimental paradigm, we demonstrate that mechanisms involving intentional action lead to intuitions of causal intransitivity, but only when these actions are norm-violating. Potential generalizations and implications for scientific theory construction are discussed.

Keywords: explanation; causal mechanisms; causal chains; transitivity; mediators

Suppose you make the observation that in a class of pupils, all girls get high grades in a gym class, while all boys get low grades. You get the impression that gender might be an important factor to explain the grades in this class: being a girl seems to be causally relevant for a pupil to get high grades. Suppose further investigations bring to light how this causal relationship is brought about: being a girl causes you to have more flexible joints, which in turn causes you to perform better in the gym class. Given this information, does being a girl still explain why someone gets high grades? Intuitively, it does: the causal relationship of gender on grades is *implemented* or *mediated* by a physiological mechanism involving agility; the mechanism information just specifies how exactly the causal influence of gender on grades is realized.

Contrast this to your intuitions towards the following alternative information about how exactly gender influences grades in the observed class: the gym teacher likes girls better than boys and gives high grades to everyone he likes. In other words, being a girl causes you to be liked by the teacher, and being liked by the teacher causes you to receive a high grade. Intuitively, gender suddenly seems less relevant in explaining the high grades. The teacher's judgments seem to be an *alternative* explanation for the grades rather than a characterization of how exactly gender exerts its causal impact on the grades.

This example illustrates that there are different possible conceptualizations of causal mechanisms. Some intermediate causes in chains are seen as *mediators* explaining how exactly the root cause brings about its effect, while other intermediate causes are seen as *alternative explanations* for the effect in question, replacing the root cause as appropriate explanation. Furthermore, the example indicates that both intuitions can be triggered for one and the same cause-effect relationship

(gender influencing grades) and with constant structural parameters (e.g., objective contingencies between the involved variables), depending only on the content of the mechanism that turns out to realize the relationship in question.

In the present paper, we will begin to investigate the psychological mechanisms that might bring about this switch in intuitions. We will first conceptualize the issue as causal transitivity problem and relate it to previous work in the field. We then turn to the question whether causal mechanisms involving intentional actions of human agents might lead to intuitions of causal intransitivity. The first experiment introduces a new paradigm designed to demonstrate the existence of the different intuitions towards the introductory example in laypeople. The second experiment tests two hypotheses as to what aspects of intentional action lead to intuitions of causal intransitivity using a different cover story. In the General Discussion, we point at implications that our results might have for transitivity intuitions outside the narrow domain of mechanisms involving intentional agents.

Transitivity in Causal Chains

Normatively, this issue can be conceptualized as the question of transitivity in causal chains. According to most classical accounts of probabilistic causality (e.g., Eells, 1991) and to the calculus of Bayesian networks (Pearl, 2000), principally, if A causes B and B causes C, then it follows that A causes C. Learning about the mechanism that implements a known causal relationship between A and C should therefore not affect the assessment of this known relationship. Technically, we can ask whether we should hold the value of B fixed when assessing the causal impact of A on C. The answer is no because B is not causally independent of A (see Rehder, 2014). For many examples, this is intuitively clear: if I want to assess whether my drinking four pints in the evening (A) causes me to have a headache on the next day (C), I should not hold fixed the amount of toxins produced in my body in the meantime (B). Causality is transitive: *A per se* is seen to be critical for C even when it is established that its causal influence is entirely mediated via B—as in the agility version of the introductory example.

However, the intuition that causality is transitive in causal chains is not always observed in the causal judgments of laypeople. Recently, Johnson and Ahn (2015) presented subjects with descriptions of numerous token causal chains (e.g., "Allison exercised for 20 min [A], then Allison became thirsty [B], then Allison drank a whole bottle of Water [C]")

and then asked them to what extent they would say that A caused B, to what extent B caused C, and to what extent A caused C. For some examples, like the one above, they found high ratings for all three causal relationships, indicating transitivity. For other token chains, however, causality was not seen as transitive. For example, in “Ned ate very spicy food [A], then Ned drank a lot of water [B], then Ned had to urinate [C]”, people rendered high causality ratings for $A \rightarrow B$ and for $B \rightarrow C$, but much lower ratings for $A \rightarrow C$. Thus, even though both links of the chain were seen as highly causal, the root cause was judged to have only a low causal impact on the final effect. This is analogous to our intuitions towards the teacher version of the introductory example: despite the fact that the teacher’s sympathy (B) is causally dependent on the pupil’s gender (A), it seems that B is seen as an alternative cause of the grades (C) rather than as a descendant of A implementing the mechanism leading from A to C.

Note that in these cases it is not denied that A causes B. The data by Johnson and Ahn (2015) indicate that people can be fully aware of this relationship, yet see B in some sense as an independent cause of C (as it can be causal for C where A is not). Thus, the judgment that A *per se* is not causal for C does not reflect a belief that there is no (indirect) causal connection between A and C. Rather, it seems to reflect the intuition that A’s causal impact on C is mediated via the “wrong” kind of mechanism. In other words, there seem to be some mechanisms which are compatible with the notion that the root cause *per se* matters for the effect, while there are other mechanisms which are incompatible with this notion.

Johnson and Ahn (2015) used token causal chains describing everyday actions of individuals and their effects. Causal transitivity varied widely across their individual items (see examples above). They showed that the extent to which such chains are causally transitive is a function of the extent to which the chains are represented in semantic memory as one single chunk (rather than as two separate relationships that are stored independently). However, the data do not allow conclusions as to which *item* properties cause a chain to be represented one way rather than the other. It thus seems unclear how their account could handle our intuitions towards the introductory example, where new mechanism knowledge is discovered for a constant unfamiliar type-level causal relationship which is unlikely to have a pre-established representation in semantic memory. Additional processes seem to be at work here.

Intentional Agents Implementing Causal Chains

Which features of a discovered mechanism determine the resulting transitivity intuitions? A salient property of the teacher mechanism in the introductory example is that it involves an intentional agent as realizer of the causal relationship in question, while the agility mechanism is part of a blind biological process. Previous research suggests that intentional actions are particularly likely to be selected as the

principal cause of terminal effects in unfolding token causal chains. Hilton, McClure, and Sutton (2009) have shown that when asked to identify the actual cause of a token event, people trace back the antecedent causal chain until they reach an intentional agent and designate his action to be the principal cause of the effect—even if the downstream causal process involves highly abnormal events that would be designated to be the principal cause in the absence of upstream intentional agency (see also Hilton & Slugoski, 1986). In other words, intentional root causes seem to make chains transitive even when they involve highly abnormal events. Accordingly, one may suspect that the reverse might also hold: finding out that a physical root cause influences its effects by affecting intentional agents’ decisions may make the chain *intransitive*. Intentional agents may be generally seen as unmoved movers that initiate causal processes rather than merely transfer external influences, producing intransitive chains and screening off the influence of the root cause from the explanandum.

However, in the materials used by Hilton et al. (2009), the intentional actions that were selected as explanations for their distal effects were usually also morally wrong or at least highly negligent. The same is true for the teacher’s grading practice in the introductory example which obviously involves morally dubious criteria. According to Hitchcock and Knobe (2009), morally abnormal actions tend to be selected as causes in common-effect structures. Rather than for their status as intentional actions, the explanations in the causal chains in Hilton et al. (2009) may have been selected for their status as especially abnormal events. In this case, intentional actions that are not norm-violating should not be seen as alternative explanations relative to the antecedent cues by which they are triggered, but as proper mediators instead. Experiment 2 is designed to differentiate between the intentionality and the abnormality hypotheses.

For our experiments, we did not employ a causal selection task for the explanation of a single mundane token event. Rather, we wanted to explore whether the conceptualization of one and the same type-level causal relationship can be differentially affected by learning that it is implemented by different kinds of causal mechanisms. In our experimental paradigm, we first teach all subjects the existence of a type-level causal relationship ($A \rightarrow C$) and assess (i) how appropriate it would be to state that A is crucial for C. We then provide different groups of subjects with different information about the mechanism implementing this relationship ($A \rightarrow B \rightarrow C$, where the content of B is varied between subjects). Afterwards we assess once again how appropriate it would now be to state (ii) that A is crucial for C and (iii) that B is crucial for C. If the second rating for A is as high as the first rating for A, this will indicate that the mechanism elicited a transitivity intuition: the fact that A causes C *via* B is compatible with the notion that A *per se* matters for C. By contrast, if the rating for A is decreased in response to learning about a specific mechanism while the rating for B is at least as high as the first rating for A, this will indicate that the chain is seen

to be intransitive: the intermediate cause is conceptualized as an *alternative explanation*, replacing A as appropriate explanation for C.

Experiment 1

In the first experiment we sought to establish that the paradigm outlined above is able to capture the intuitive differences displayed in the introductory example.

Participants

The experiment was conducted as an online study. A total of 171 subjects from the UK completed the experiment, 32 of which were excluded from the statistical analyses because they failed in a simple attention check question that we asked at the end of the experiment. The average age of all included subjects ($N = 139$, 93 women) was 38 years ($SD = 8.62$).

Design, Materials, and Procedure

We constructed a complete 2 (Mechanism: Physiology vs. Teacher) \times 2 (Contingency: Deterministic vs. Probabilistic) \times 2 (Balance: Boys with high grades vs. Girls with high grades) between-subjects design. Subjects in all conditions were asked to take the perspective of a scientist investigating the relationship between pupils' gender (A) and their grades in a physical education class (C). In a first learning phase they received data of a class of ten pupils (five boys and five girls) in tabular form which showed each pupil's gender (A vs. $\neg A$) and whether he or she got a high grade (C) or a low grade ($\neg C$). Whether being a boy or a girl was designated as A was counterbalanced with the Balance factor. When boys had high grades, the grades in question were for a course in athletics; when girls had high grades, they were for a course in gymnastics. Half of the subjects learned that there was a deterministic relationship between gender and grades, for the other half this relationship was probabilistic (one exception for each gender). This contingency manipulation was included to explore whether intransitivity intuitions can be elicited for both deterministic and probabilistic causal relationships. After having studied this table, participants were asked to indicate on an 11-point scale (ranging from 0 to 10) how appropriate they found the following sentence to describe the concrete observations they have just made: "The gender of a pupil is crucial for this pupil's grade in athletics/gymnastics" (we call this *appropriateness rating* $[A \rightarrow C]_{pre}$ because it is measured prior to the introduction of mechanism information). At this point, we expected that subjects would have formed the impression that, within the observed sample, gender influences grades ($A \rightarrow C$), leading them to render unanimously high appropriateness ratings.

The crucial mechanism manipulation was introduced in a second learning phase. Subjects were told that they would have come up with a hypothesis about the underlying mechanism. Half of them (Mechanism: Physiology) were told that they suspected boys to develop higher muscularity than girls which in turn would lead them to receive higher grades in athletics. (In the other Balance condition, girls were sus-

pected to develop higher agility than boys which in turn would lead them to receive higher grades in gymnastics.) The other half (Mechanism: Teacher) was told that they suspected the teacher to like boys better than girls (and vice versa for the other Balance condition), leading boys (girls) to receive higher grades. In all conditions, subjects read that they went back to collect additional data from the same class corresponding to their hypothesis. These data were then presented in a new version of the table which now included an additional column representing each pupil's value on the suspected mediating variable (B [high] vs. $\neg B$ [low]; in both Contingency conditions, this new variable deterministically predicted the grades). After having studied this extended table, subjects were again asked to indicate how appropriate it would be to state that A is crucial for C ($[A \rightarrow C]_{post}$), and also how appropriate it would be to state that B is crucial for C ($[B \rightarrow C]_{post}$) using the same question format (B being described as "the muscularity of a pupil/the agility of a pupil/the teacher's sympathy for a pupil", depending on condition). We expected continuously high ratings for $(A \rightarrow C)_{post}$ in the Physiology condition and a drop in ratings for $(A \rightarrow C)_{post}$ in the Teacher condition.

Afterwards, we assessed contingency estimates for all three relationships from memory (i.e., $P[C|A]$ vs. $P[C|\neg A]$, $P[B|A]$ vs. $P[B|\neg A]$, and $P[C|B]$ vs. $P[C|\neg B]$), assessed on six separate 11-point scales ranging from 0 (impossible) to 100 (certain) to see if subjects in both Mechanism conditions based their judgments on the same impression of the objective probabilities. In the Probabilistic conditions, we furthermore asked them to indicate the conditional dependence of C on A given constant values of B (i.e., $P[C|A \wedge B]$ vs. $P[C|\neg A \wedge B]$, and $P[C|A \wedge \neg B]$ vs. $P[C|\neg A \wedge \neg B]$) to see if they understood that, in the observed sample, C was independent of A when B was held fixed (regardless of the Mechanism condition). Finally, we wanted to know how plausible they found each of the described causal relationships ($A \rightarrow C$, $A \rightarrow B$, and $B \rightarrow C$) according to their prior real world knowledge, regardless of the fictional data from the experiment.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 1. We conducted a global 2 (Mechanism: Physiology vs. Teacher) \times 2 (Contingency: Deterministic vs. Probabilistic) \times 2 (Balance: Boys high vs. Girls high) \times 3 (Rating: $(A \rightarrow C)_{pre}$ vs. $(A \rightarrow C)_{post}$ vs. $(B \rightarrow C)_{post}$, within-subject) mixed ANOVA. Since there was neither a main effect of Balance, $F_{1,131} < 1$, nor any significant interaction effect including Balance, largest $F_{2,262} = 2.19$, data in Figure 1 are averaged across this factor. There was a main effect of Rating, $F_{2,262} = 22.00$, $p < .001$, $\eta_p^2 = .14$, and a significant interaction of Rating \times Contingency, $F_{2,262} = 5.99$, $p = .003$, $\eta_p^2 = .04$. There was also a trend for an interaction of Rating \times Mechanism, $F_{2,262} = 2.87$, $p = .06$, $\eta_p^2 = .02$.

The main prediction that we made was that we would see a distinct drop of appropriateness ratings for $(A \rightarrow C)_{post}$ relative to $(A \rightarrow C)_{pre}$ selectively within the Teacher conditions.

A planned contrast testing whether this difference was larger in the Teacher compared to the Physiology condition confirmed this prediction, $t_{131} = 2.97$, $p < .05$, $r = .25$. Furthermore, the ratings for $(B \rightarrow C)_{\text{post}}$ were at least as high as the ratings for $(A \rightarrow C)_{\text{pre}}$ in all conditions (see Figure 1). The increase in appropriateness ratings for $(B \rightarrow C)_{\text{post}}$ in the Probabilistic conditions can be explained by the fact that B is a deterministic predictor for C while A is not (i.e., the probabilistic element lies in $A \rightarrow B$).

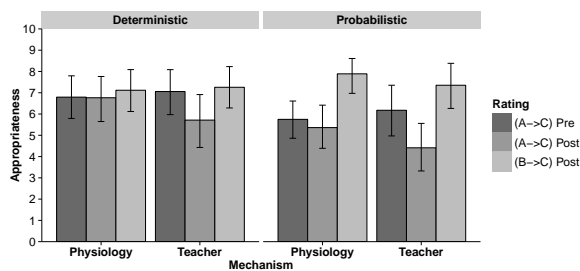


Figure 1: Group means (Errorbars = 95% CI) of appropriateness ratings in Experiment 1.

An analysis of the subjects' contingency estimates showed that, globally, subjects represented the contingencies contained in the learning data quite adequately, apart from generally underestimating the strength of the deterministic relationships. Crucially, this pattern was similar in both the Physiology and the Teacher condition. Furthermore, participants in the Probabilistic conditions recognized that A and C were conditionally independent given constant levels of B ($P[C|A \wedge B] \approx P[C|\neg A \wedge B]$ and $P[C|A \wedge \neg B] \approx P[C|\neg A \wedge \neg B]$). This pattern was also consistent in both Mechanism conditions. The pattern of contingency estimates therefore cannot account for the differences we observed for the appropriateness ratings between the Mechanism conditions. The same holds for the plausibility ratings: the relationships $A \rightarrow B$ and $B \rightarrow C$ were rated to be equally plausible for both mechanisms.

Discussion

The results of this experiment show that acquiring different mechanism knowledge can differentially affect the conceptualization of a given type-level causal relationship. When effects of gender on grades in a gym class were mediated via a physiological process, this was seen as compatible with the notion that gender *per se* matters for the grades. This was not the case when the relationship was mediated via the teacher's personal preference for the pupil. In this case, subjects revised their earlier judgment that gender was critical for the grade and attributed the grades to the teacher's sympathy instead, even though they were aware that the teacher's preference was causally affected by the pupils' gender. This result cannot be explained by differences in encoded contingencies or prior plausibility intuitions for the different mechanisms.

Experiment 2

In Experiment 2, we attempted to replicate the phenomenon using a more artificial cover story. This story allowed us to differentiate between the intentionality and the abnormality hypotheses developed in the theory section: is the root cause always screened off when the mechanism involves an intentional agent, or only when this agent's intentions are morally dubious?

Participants

The experiment was conducted as an online study. We recruited 232 subjects from the UK, 31 of which failed in the attention test and were excluded from all analyses. The average age of all included subjects ($N = 201$, 118 women) was 36 years ($SD = 8.14$).

Design, Materials, and Procedure

We used the same experimental paradigm as in Experiment 1. This time, we implemented four between-subjects conditions describing different mechanisms underlying the same causal relationship. Subjects in all conditions were asked to imagine they were at a funfair where they were observing a swing tossing passengers around. They read that sometimes after a passenger had entered the swing a red flashlight came on, whereupon the passenger had to leave the swing without having taken a ride. Participants read that they would have gained the impression that the flashlight (C) came on more frequently for corpulent passengers (A). Upon studying the learning data about a deterministic relationship between A and C across ten observed passengers (the first five of them corpulent, the last five not corpulent), they rendered their $(A \rightarrow C)_{\text{pre}}$ appropriateness rating.

We then told subjects in the different conditions that they would have come up with one of four hypotheses about the underlying mechanism. In the first condition (Scale) they were told that they suspected the flashlight to be part of a safety mechanism. They believed a scale to be built into the swing which causes the flashlight to come on whenever the loading exceeds a critical threshold. This condition was intended to provide a baseline for unequivocal judgments of transitivity, analogous to the Physiology condition in Experiment 1. In the second condition (Accurate Operator) the subjects' hypothesis was that the operator intends to guarantee the safety of his costumers, and that whenever he judges a passenger to be too corpulent for his swing he activates the flashlight. This condition was intended to provide a mechanism involving an intentional agent but otherwise being closely matched to the Scale condition. According to the intentionality hypothesis, this chain should nonetheless be seen as intransitive, while it should be seen as transitive according to the abnormality hypothesis since the operator does not violate a norm. The mechanism hypothesis of the third condition (Biased Operator) was that the operator does not like corpulent people and enjoys embarrassing them, and that whenever he judges a passenger to be too corpulent for his taste, he activates the flashlight. This condition was intended as providing

an intentional mechanism analogous to the Teacher condition in Experiment 1 which should elicit judgments of intransitivity according to both the intentionality and the abnormality hypothesis. In the last condition (Central Computer), the subjects' hypothesis was that a central computer supervising the electricity supply detects irregularities at unpredictable times, and then generates an electronic signal that causes the flashlight to come on. The co-occurrence of corpulence and flashlight in the observed sample would have resulted from mere coincidence. This condition was intended to create a structure in which A turned out to be actually causally irrelevant for C. B (the central computer) caused the effect, and the relationship between A and C in the sample was merely due to a coincidental correlation between A and B in the observed sample. This condition thus provides a baseline for an unequivocal alternative explanation. The crucial question is where intuitions towards the mechanisms involving intentional operators fall within the space that is spanned between Scale (a clear mediator) and Central Computer (a clear alternative explanation).

In all conditions, subjects read that they went over to the operator and asked him for information concerning the mechanism. The operator confirmed their hypothesis in all conditions and told them for the last ten passengers whether or not the scale had detected a threat to the passenger/he had detected a threat to the passenger/he had not liked the passenger/the central computer had detected irregularities in the electricity supply (depending on condition). After having studied the corresponding extended table in which all three relationships were deterministic, subjects again rendered their appropriateness ratings for $(A \rightarrow C)_{\text{post}}$ and $(B \rightarrow C)_{\text{post}}$. Finally, subjects indicated their contingency estimates for all three relationships from memory. Plausibility ratings were not collected.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 2. We conducted a global 4 (Condition: Scale vs. Accurate Operator vs. Biased Operator vs. Central Computer) \times 3 (Relationship: $(A \rightarrow C)_{\text{pre}}$ vs. $(A \rightarrow C)_{\text{post}}$ vs. $(B \rightarrow C)_{\text{post}}$, within-subject) mixed ANOVA. There was no main effect of Condition, $F_{3, 197} = 1.81$, but a main effect of Relationship, $F_{2, 394} = 7.48$, $p < .001$, $\eta_p^2 = .04$. The global Relationship \times Condition interaction was not significant, $F_{6, 394} = 1.74$, $p = .11$, $\eta_p^2 = .03$. However, planned contrasts revealed that the decrease in appropriateness ratings from $(A \rightarrow C)_{\text{pre}}$ to $(A \rightarrow C)_{\text{post}}$ did not differ between Scale and Accurate Operator, $t_{197} < 1$, but was larger compared to Scale both in the Biased Operator condition, $t_{197} = 2.41$, $p < .05$, $r = .17$, and in the Central Computer condition, $t_{197} = 2.81$, $p < .01$, $r = .20$. Furthermore, the decrease was as large in Biased Operator as in Central Computer, $t_{197} < 1$. At the same time, $(B \rightarrow C)_{\text{post}}$ was not smaller than $(A \rightarrow C)_{\text{pre}}$ in any of the conditions, largest $t_{197} = 1.48$.

The contingency estimates were similar across all four mechanism conditions for all three causal relationships

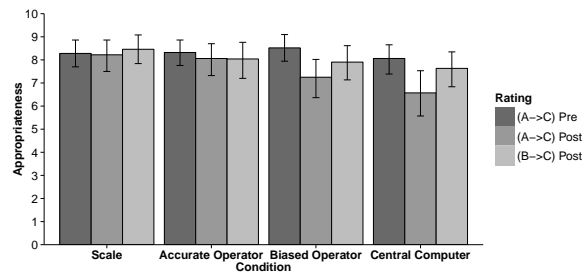


Figure 2: Group means (Errorbars = 95% CI) of appropriateness ratings in Experiment 2.

($A \rightarrow B$, $B \rightarrow C$, $A \rightarrow C$). The only exception was that ΔP for $A \rightarrow B$ tended to be higher in the Scale condition than in the other three conditions, a difference that was significant in the comparison with Biased Operator, $t_{197} = 2.28$, $p < .05$. This indicates that the contingency between corpulence and the biased operator's judgment was perceived to be weaker than the contingency between corpulence and threat detection by the scale, despite identical learning data. While this difference may add to the decrease in $(A \rightarrow C)_{\text{post}}$ in the Biased Operator condition, the overall pattern of contingency estimates cannot fully account for the overall pattern of appropriateness ratings.

Discussion

The results of Experiment 2 demonstrate that involving an intentional agent is not a sufficient property for a mechanism to elicit judgments of intransitivity. Agents intending to accurately transfer an objective signal from A to C seem to be conceptualized akin to a mechanical process serving the same function. However, if the same relationship is dependent on an agent's highly idiosyncratic (and morally dubious) preference structure, the physical root cause is screened off from the explanandum to the same extent as if it was merely a coincidental, causally irrelevant confound.

General Discussion

In two experiments, we have demonstrated that the conceptualization of one and the same established type-level causal relationship can be differentially affected when knowledge about different causal mechanisms is acquired. Some mechanisms (e.g., physiological processes) are compatible with the notion that the root cause *per se* matters for the effect, while others (e.g., biased judges) provide an *alternative* explanation, leading the root cause to be seen as a less appropriate explanation for the effect. Our data indicate that these differences are not brought about by different causal strength assessments, nor by different plausibility intuitions. Also, involving an intentional agent does not constitute a sufficient condition for a mechanism to elicit intuitions of intransitivity. In the contexts we investigated, this intuition seems to depend crucially on the intentional agent being morally abnormal.

This latter finding may indicate that a more general psychological mechanism may underlie our findings which might make our framework applicable to intransitivity intuitions outside the narrow scope of mechanism involving human agents. Immoral human agents may merely be a very salient instance of abnormal mechanisms more generally. Recently, Kominsky, Phillips, Gerstenberg, Lagnado, and Knobe (2015) have argued that moral and statistical abnormality of potential causes function similarly in eliciting counterfactual possibilities that in turn have similar downstream effects of the assessment of other causes in the same common-effect network. It is possible that similar generalizations hold in our case. In the Physiology condition from Experiment 1, the mechanism is implemented in every single pupil in a lawlike manner. Thus, it can be assumed that an $A \rightarrow C$ relationship brought about by this mechanism will be stable across most perturbations of the entity implementing the mechanism in each token case, both within and beyond the observed sample. Contrast this with the teacher example: the observed type-level $A \rightarrow C$ relationship is entirely dependent on this particular teacher implementing the mechanism in the observed sample. Replacing the teacher with pretty much any colleague would presumably make the $A \rightarrow C$ relationship disappear. Even though A is doubtlessly an indirect cause of C in the observed sample, explaining C in terms of A might feel inadequate because the relationship can be expected to be highly sensitive to minor perturbations of the boundary conditions provided by the particular teacher implementing the mechanism in the observed sample (see also Garfinkel, 1981).

So far, these are only speculations as to how far our findings may generalize which still need to be empirically tested with materials outside the domain of human agency. In case of success, this account may be applicable not only to aspects of everyday causal cognition, but even to psychological processes underlying scientific theory construction. Whenever a scientific theory is about a causal chain structure (e.g., process X influences process Y, which in turn leads to effect Z), the issue discussed in this paper arises: if the researcher wants to assess whether process X *per se* explains outcome Z, should she hold process Y constant? If she conceives of process Y as a mediator, the answer is definitely no. But there might be cases in which, despite the underlying chain structure, she conceives of process Y as an *alternative explanation* of effect Z. Sometimes, it does not seem clear which of these conceptualizations is more adequate—yet, the decision for one of them will crucially shape the methodology and conclusions of the subsequent research (e.g., decisions about whether process Y is to be controlled or to be left free to covary with X).

Acknowledgments

We thank Jana Samland, Jonathan Kominsky, Joshua Knobe, and Michael Waldmann for helpful comments.

References

- Eells, E. (1991). *Probabilistic causality* (Vol. 1). Cambridge: Cambridge University Press.
- Garfinkel, A. (1981). *Forms of explanation: Rethinking the questions in social theory*. New Haven: Yale University Press.
- Hilton, D. J., McClure, J., & Sutton, R. M. (2009). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology, 40*, 383–400.
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review, 93*, 75.
- Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy, 106*, 587–612.
- Johnson, S. G., & Ahn, W.-K. (2015). Causal networks or causal islands? the representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*.
- Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition, 137*, 196–209.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.
- Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*, 54–107.

Appendix I

Nagel and Stephan (2016)

Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 806 – 812). Austin, TX: Cognitive Science Society.

Explanations in Causal Chains: Selecting Distal Causes Requires Exportable Mechanisms

Jonas Nagel (jnagel1@uni-goettingen.de)
Simon Stephan (sstepha1@uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

When A causes B and B causes C, under what conditions is A a good explanation for the occurrence of C? We propose that distal causes are only perceived to be explanatory if the causal mechanism is insensitive to inessential variations of boundary conditions. In two experiments, subjects first observed deterministic $A \rightarrow B \rightarrow C$ relationships in a single exemplar of an unknown kind. They judged A to be crucial for C by default. However, when they subsequently learned that the causal mechanism fails to generate the $A \rightarrow C$ dependency in other exemplars of the same kind, subjects devalued A as a crucial explanation for C even within the first exemplar. We relate these findings to the idea that good explanations pick out portable dependency relations, and that sensitive causes fail to meet this requirement.

Keywords: explanation; causal mechanisms; causal chains; sensitivity; portability

Introduction

Causal relationships are implemented by causal mechanisms (Machamer, Darden, & Craver, 2000). If we say that one event (A) causes another event (C), we generally assume that there is some process leading from A to C that can principally be discovered, even if we do not yet know what the actual mechanism is. Accordingly, the causal arrow in notations such as $A \rightarrow C$ is sometimes interpreted as a “mechanism placeholder” (Pearl, 2000).

When mechanism knowledge about a particular causal relationship (e.g., $A \rightarrow C$) is made explicit, the resulting causal model takes the form of a causal chain (e.g., $A \rightarrow B \rightarrow C$, where B is an intermediate cause implementing the mechanism). The current research asks how such integration of mechanism knowledge affects people’s conceptualization of the original causal relationship. More specifically, the question is which properties of mechanism B can affect people’s impression of the importance of A for explaining C.

Intuitively, there are two ways to interpret the causal role of B in causal chains. First, B could be seen as a *mediator* implementing the causal influence of A on C. Under this reading, A continues to be seen as explanatory for C, even though its causal influence is completely mediated via B. Second, it could be seen as an *alternative explanation* for the occurrence of C, screening off the influence of A on C. The fact that, if we know the value of B, A adds nothing to explaining C, provides a reason to devalue A as appropriate explanation of C under this interpretation.

Nagel and Stephan (2015) showed that both interpretations can arise when subjects learn that different mechanisms implement one and the same type-level causal relationship. They had their subjects learn a strong dependency of grades

in a gym class (C) on the pupils’ gender (A) from fictional covariation data. Subjects indicated strong agreement with the claim that, within the observed sample, a pupil’s gender was crucial for his or her grade. In a second learning phase, half of the subjects learned that the $A \rightarrow C$ relationship was mediated by a genetically determined physiological process (B_1), while the other half learned that it was mediated by the gender preferences of the teacher (B_2). Afterwards, subjects in the physiological mechanism condition continued to endorse the statement that a pupil’s gender was crucial for this pupil’s grade (indicating that B_1 was interpreted as a mediator of the original $A \rightarrow C$ relationship), while subjects in the teacher mechanism condition devalued gender as crucial explanation for the grades (indicating that B_2 was interpreted as an alternative explanation for C).

Both conditions were equivalent in terms of objective causal structure and observed dependency patterns, so it seems the difference in interpretation results from some aspect of the manifold content-related differences between both contrasted mechanisms. One salient hypothesis is that intentional agents, like the teacher, might be seen as initiators of causal sequences (unmoved movers) and therefore always screen off the influence of upstream physical preconditions of their actions from downstream effects. Blind physical processes like genetics, by contrast, may not have this quality and may thus be regarded as mere mediators of the influence of upstream root causes. In a second experiment, Nagel and Stephan (2015) ruled out this hypothesis. They presented their subjects with a scenario in which a physical signal (A) was picked up by a human agent who deliberately reacted to the signal (B) to produce a final outcome (C). Subjects repeatedly observed this deterministic causal chain in all conditions. Half of the participants learned that the agent had a benign motivation in implementing the $A \rightarrow C$ chain, while the other half learned that he had a malevolent motivation. Afterwards, they were asked to what extent it was appropriate to state that the original signal (A) vs. the agent’s reaction (B) was crucial for the occurrence of the outcome (C). It turned out that in case of the benign agent, both the distal signal and the proximal action were judged to be equally crucial for the occurrence of C. The distal physical cause thus retained its explanatory power and was seen to bring about the outcome *by means of* human agency. By contrast, if the agent had a morally dubious motive, the distal cause was devalued as explanation despite a perfect dependency relation with the outcome in the observed sample; proximal human agency served

as *alternative explanation* for the outcome instead. This finding strongly suggests that it is not an agent's intentionality per se that leads to devaluation of upstream causes as explanations for downstream effects, but rather some other property that is related to the motivation of the agent. In the remainder of this paper, we will outline and test the hypothesis that this property does not reside exclusively in moral qualities of intentional agents, but quite generally reflects inferences about whether the mechanism can be expected to generalize to other, inessentially different contexts.

Sensitivity and Explanatory Relevance

Woodward (2006) investigated the human practice of making causal claims. He noted that causal claims require not only that the effect be counterfactually dependent on the cause, but also that this counterfactual dependence continue to hold under varied boundary conditions. Causal relationships that do not fulfill this second requirement are called *sensitive*, and Woodward (2006) argues that sensitive causal relationships are regarded as deficient despite a strong dependence relationship under the conditions in which they do obtain. Good causes are those that not only bring about their effects in the narrow context of actually observed circumstances, but would continue to do so in different contexts. The requirement for good causes to be insensitive resonates with philosophical and psychological accounts of explanation. Many theorists have argued that good explanations tend to pick out factors that are generalizable beyond the concrete set of observations that presently is to be explained (Garfinkel, 1981; Hitchcock, 2012; Lombrozo & Carey, 2006). This ensures that the generated explanation will be useful for making predictions and for planning interventions in future similar situations (Lombrozo, 2010).

One reason for high sensitivity of causal relationships is that the mediating mechanism works reliably only under quite specific boundary conditions, but is easily disturbed in other, similar situations. We propose that whenever people find out that an observed causal relationship is implemented by a mechanism that is highly sensitive in this sense, they devalue the distal cause as explanation for the terminal effect. To illustrate, consider again the scenarios used by Nagel and Stephan (2015). If the influence of gender on grades is mediated by a genetically determined physiological mechanism, this implies that the relationship will continue to hold in future observations with different samples of pupils, which makes the $A \rightarrow C$ relationship insensitive. The teacher mechanism, by contrast, implies that this relationship depends on the presence of highly peculiar boundary conditions which will rarely be met in other, similar situations (as most other teachers, hopefully, will not exhibit the same bias). This makes the observed $A \rightarrow C$ relationship highly sensitive—it will break down as soon as we leave the narrow context of the class that was actually observed in the sample. It is recognized that gender will not *generally* influence grades and is therefore considered a poor explanation for the grades *even*

within the observed sample.

The mechanisms compared by Nagel and Stephan (2015) differed on many dimensions other than sensitivity, including intentional agency and moral abnormality. Our goal in the present experiments is to isolate the sensitivity of the mediating mechanism. We created new experimental material from the domains of biology and physics in which we manipulated a given mechanism's sensitivity purely in statistical terms. We first presented subjects with a single exemplar of an unknown natural kind or artifact and let them discover a deterministic $A \rightarrow B \rightarrow C$ chain within this entity. In a subsequent learning phase, we showed subjects the same exemplar again, but this time in the company of several other exemplars of the same kind with identical appearance. One half of the subjects saw that the new exemplars behaved just like the first exemplar in terms of the $A \rightarrow B \rightarrow C$ dependency pattern. The other half saw that only the first exemplar once again showed the same dependency pattern, while in all other exemplars the presence of A failed to lead to the presence of B (and, hence, C). Subjects in both conditions were then asked how appropriate it was to say that A was crucial for the presence of C *within the first exemplar only* which had constantly displayed perfect dependency relations in both conditions. We predicted reduced appropriateness ratings in the condition in which the dependence of C on A did not generalize to other exemplars of the same kind.

In the first experiment, we tested and confirmed these predictions in the domain of biology. In the second experiment, we replicated the findings in the domain of artifacts and additionally controlled for the relative complexity of the potentially explanatory variables A and B.

Experiment 1

Participants

The experiment was conducted as an online study. A total of 150 subjects were recruited from a panel (www.pureprofile.com), 44 of which (29%) were removed prior to the analyses because they did not complete the survey or failed to solve a simple attention check question at the end. The mean age of all included subjects ($N = 106$, 80 women) was 37 years ($SD = 8.16$). Included subjects received a payment of £ 6 per hour.

Design, Materials, and Procedure

Subjects were randomly allocated to the conditions (Sensitivity: Sensitive vs. Insensitive). They were asked to take the perspective of a marine biologist who discovered a single exemplar of deep sea fish and called it "Fish #1". Their task was to test whether noise (A) leads to activity in the brain of the fish (B) and finally to an illumination of the fish's antenna (C). We then presented our subjects an animation of Fish #1 (see Figure 1). When participants pressed the "Play" button, they saw sound waves coming out of the speaker. About half a second later, the brain activity device's monitor displayed a flickering amplitude moving across the screen. Finally, about

one second later, the fish’s flash bulb turned from blue to yellow. If participants hit the stop button, all variables returned to their initial state. Subjects could (de-)activate the loudspeaker as often as they wished.

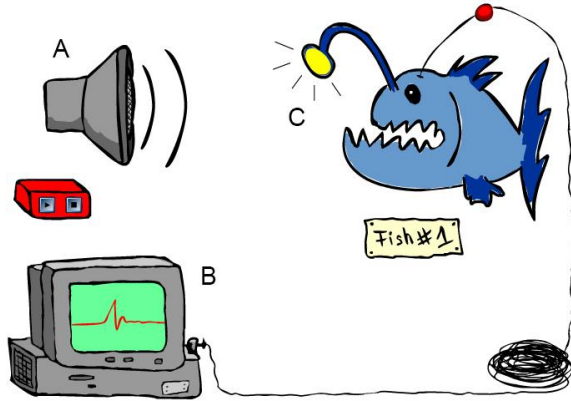


Figure 1: Screenshot of the animation used in the first learning phase of Experiment 1 while all three variables are active. Letters A, B, and C were not shown to participants.

When participants felt they had learned enough about the relationships, they proceeded to the first question screen on which they were asked how appropriate the following statements were for describing the observations they had just made of Fish #1. The first statement was “The presence of sound waves is crucial for Fish #1’s antenna to lighten up” (appropriateness rating $[A \rightarrow C]_{pre}$), and the second statement was “Activity of the brain area is crucial for Fish #1’s antenna to lighten up” (appropriateness rating $[B \rightarrow C]_{pre}$). We used two independent rating scales to allow participants to judge both causes as equally explanatory, while at the same inviting them to see both statements as contrastive alternatives by using the word “crucial”. Participants provided their judgments on 11 point rating scales ranging from “0 = not at all appropriate” to “10 = very appropriate” for each statement. We expected equally high ratings for both statements, indicating that brain activity is seen as a mediator and sound waves are seen to be explanatorily relevant.

The experimental manipulation was applied after subjects had given their baseline ratings. All participants read that they had caught nine additional exemplars of the same kind of fish. On the next screen, they saw a large animation showing all ten fish (consecutively labelled Fish #1 to Fish #10), each in the same set-up as shown in Figure 1. Below the ten fish, there was a device with a play- and stop-button which they could use to (de-)activate all ten loudspeakers simultaneously. In both conditions, Fish #1 again reacted exactly as in the first learning phase. The crucial difference between both conditions was the behavior of the additional nine fish exemplars. In the insensitive condition, all other fish behaved exactly like

Fish #1, while in the sensitive condition, none of the other fish reacted to the activation of the loudspeaker with brain activity or antenna lightning. This manipulation was intended to confirm in the insensitive condition the assumed prior expectation that the $A \rightarrow C$ dependence is exportable from Fish #1 to the whole kind, while subjects in the sensitive condition should conclude that the $A \rightarrow C$ dependence is not exportable beyond the narrow context of Fish #1.

After having made these additional observations, subjects were asked to reconsider the results of their previous experiment with Fish #1 only and to answer the same two questions again in light of their new knowledge about the whole swarm of fish. The appropriateness ratings $(A \rightarrow C)_{post}$ and $(B \rightarrow C)_{post}$ were measured exactly as the baseline ratings described above. Our central prediction for the sensitive condition was that the $(A \rightarrow C)_{post}$ ratings should drop considerably because the $A \rightarrow C$ dependency is not exportable beyond the context of Fish #1. The $(B \rightarrow C)_{post}$ appropriateness ratings, by contrast, should not be reduced by the swarm information. Brain activity remains a good predictor of antenna flashing across the whole swarm. In the insensitive condition, of course, neither of the ratings should be affected by the swarm data.

Finally, we wanted to make sure that participants encoded the contingencies between the variables accurately both within Fish #1 and across the whole swarm of fish. On a first screen, participants were prompted to recall what they had learned about Fish #1 and were asked six conditional probability questions concerning Fish #1 only. For example, $P(C|A)$ was assessed with the question “How likely is it for Fish #1’s antenna to lighten up given that sound waves are present?” and an 11 points rating scale ranging from 0 (impossible) to 100 (certain). Analogous questions were asked for $P(C|\neg A)$, $P(B|A)$, $P(B|\neg A)$, $P(C|B)$, and $P(C|\neg B)$. On a second screen, the same six questions were repeated with the whole swarm as reference class. These twelve estimates were used to compute contingency estimates (ΔP) for each of the three causal relationships both at the exemplar level and at the swarm level. Equally high contingency estimates at the exemplar level for $A \rightarrow C$ and $B \rightarrow C$ would demonstrate that the expected effects on the appropriateness ratings would not be due to different dependency assumptions within the exemplar, but rather due to sensitivity of the $A \rightarrow C$ relationship across the whole kind.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 2a. We conducted a three-way 2 (Sensitivity: Sensitive vs. Insensitive, between-subjects) \times 2 (Relationship: $A \rightarrow C$ vs. $B \rightarrow C$, within-subject) \times 2 (Rating Position: Pre vs. Post, within-subject) mixed ANOVA. We obtained a significant two-way Sensitivity \times Position interaction, $F_{1,104} = 8.61$, $p < .01$, $\eta_g^2 = .014$, indicating that the swarm information affected appropriateness ratings in the sensitive condition more than in the insensitive condition. More importantly, this interaction was qualified by a

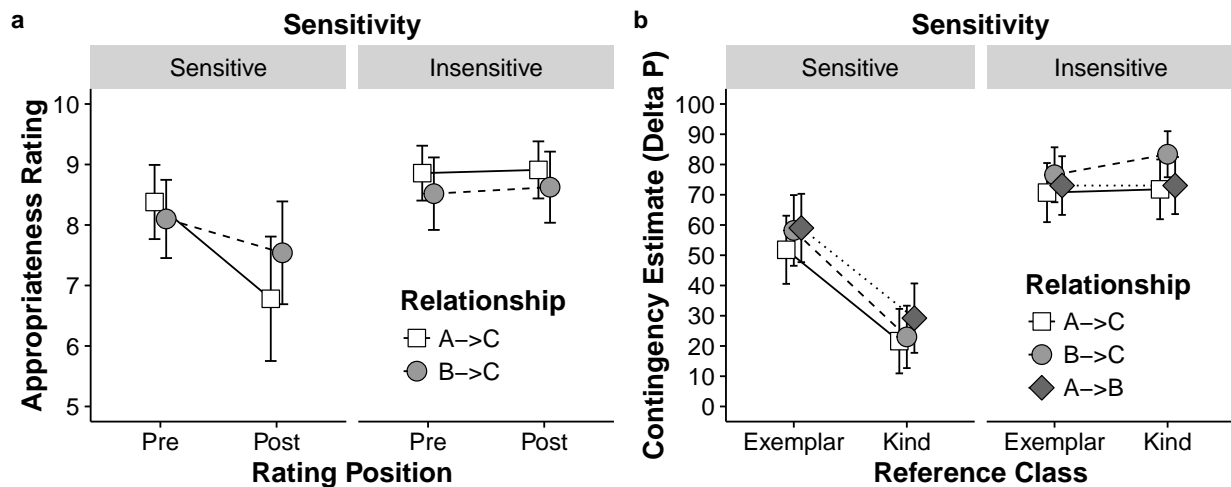


Figure 2: Group means (error bars = 95% CI) of appropriateness ratings (a) and contingency estimates (b) in Experiment 1.

marginally significant three-way Condition \times Position \times Relationship interaction, $F_{1,104} = 3.64$, $p < .06$, $\eta_g^2 = .003$, indicating that the selective decrease of ratings from pre to post in the sensitive condition was more pronounced for the $A \rightarrow C$ relationship than for the $B \rightarrow C$ relationship, as predicted by our account.

To assess subjects' contingency estimates for the three relationships, we calculated each subject's ΔP estimate for each relationship both within Fish #1 and across the whole swarm of fish. For example, the contingency estimate for the $A \rightarrow C$ relationship within Fish #1, $\Delta P(A \rightarrow C)_{\text{Exemplar}}$, was calculated by subtracting each subject's $P(C|\neg A)_{\text{Exemplar}}$ rating from the same subject's $P(C|A)_{\text{Exemplar}}$ rating. The contingency estimates across the whole swarm were calculated analogously using the conditional probability judgments for the whole swarm. The descriptive results are summarized in Figure 2b. Most importantly, the $\Delta P(A \rightarrow C)_{\text{Exemplar}}$ estimates in the sensitive condition were not lower than the $\Delta P(B \rightarrow C)_{\text{Exemplar}}$ estimates. This finding rules out the alternative explanation that the selective drop in the $(A \rightarrow C)_{\text{post}}$ appropriateness ratings in the sensitive condition results from selectively decreased dependency estimations within Fish #1 for this particular relationship.¹

¹The surprisingly low $\Delta P(B \rightarrow C)_{\text{Kind}}$ estimate in the sensitive condition resulted from very low $P(C|B)_{\text{Kind}}$ ratings. The observation that in Fish #1 (the only exemplar in which brain activity was ever recorded) brain activity reliably led to antenna illumination did not suffice to make subjects generalize this relationship across the whole kind. Hesitance to generalize from sparse data, however, is different from gathering positive evidence for sensitivity. The finding that subjects continued to regard brain activity as the crucial explanatory factor within Fish #1 despite low kind-general contingency estimates thus does not directly disprove our hypothesis, but is certainly a finding that needs further investigation.

Discussion

In this experiment, we have demonstrated that sensitive mechanisms reduce the explanatory relevance of distal causes in causal chains. Our subjects first learned that, within the narrow context of a single exemplar of a biological kind, cause A deterministically produced effect C, and that this causal relationship was always mediated by mechanism B. At this point, they interpreted the B to be a mediator of the observed $A \rightarrow C$ relationship and correspondingly found that A and B were equally crucial for the occurrence of C. However, if they subsequently found out that cause A failed to activate mechanism B in all other exemplars of the same kind, this affected their representation of the perfect dependency relation within the initially observed exemplar. Even within this exemplar, they now judged A to be less crucial for the occurrence of C than the more proximal B, indicating that A became deficient as explanation for C, despite the perfect $A \rightarrow C$ dependency relation that was observed throughout for this exemplar.

Experiment 2

In the second experiment, our main goal was to replicate the findings from Experiment 1 in the domain of artifacts. Furthermore, we made the contents of causes A and B more similar to each other in order to rule out alternative explanations for their differential treatment in the post-ratings. In Experiment 1, A was a simple, physical variable external to the system under study, while B was a complex, physiological variable internal to the system. In Experiment 2, we used only internal, physical variables that varied in complexity. We counterbalanced the assignment of the simple and the complex variable to the positions of distal cause A and proximal cause B. We expected the same pattern of results as in Experiment 1 under both assignments, showing reduced explanatory relevance of the distal cause results from mechanism sensitiv-

ity per se, regardless of its surface characteristics.

Participants

We recruited and compensated 331 subjects as in Experiment 1. 115 (35%) were removed prior to analysis according to the same criteria as above. The mean age of all included subjects ($N = 216$, 109 women) was 40 years ($SD = 8.36$).

Design, Materials, and Procedure

Subjects were randomly allocated to one of four conditions that resulted from a 2 (Sensitivity: Sensitive vs. Insensitive) \times 2 (Complexity of Mechanism: Complex vs. Simple) between-subjects design. They encountered an exemplar of an unknown machine, “Machine #1”, with three visible devices: a single rack wheel, a complex system of rack wheels, and a fan (see Figure 3). Subjects in the complex mechanism condition saw the three devices in the arrangement shown in Figure 3, while for subjects in the simple mechanism condition, the positions of the single wheel and the complex system of wheels were reversed. The procedure was analogous to Experiment 1. In a first learning phase subjects set the leftmost device in motion and observed that this was followed by movement of the device in the middle, which was in turn followed by movement of the fan on the right. Switching off the leftmost device resulted in subsequent inertia of all variables. On the next screen, we assessed appropriateness ratings ($A \rightarrow C$)_{pre} and ($B \rightarrow C$)_{pre} as in Experiment 1. In the second learning phase, subjects were shown Machine #1 again together with five additional machines with identical surface features, consecutively labelled “Machine #2” to “Machine #6”. They could separately intervene on each machine’s leftmost device as often as they wished. In the Insensitive condition, all six machines behaved just as Machine #1 in the first learning phase. In the Sensitive condition, Machine #1 also worked just as before, but in none of the additional machines did the device in the middle or the fan ever turn on. On the following screens, subjects again indicated their ($A \rightarrow C$)_{post} and ($B \rightarrow C$)_{post} appropriateness ratings as well as their exemplar-specific and kind-general conditional probability judgments analogous to Experiment 1.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 4. We conducted a four-way 2 (Sensitivity: Sensitive vs. Insensitive, between-subjects) \times 2 (Relationship: $A \rightarrow C$ vs. $B \rightarrow C$, within-subject) \times 2 (Rating Position: Pre vs. Post, within-subject) \times 2 (Complexity of Mechanism: High vs. Low, between-subjects) mixed ANOVA. We again obtained a significant two-way Sensitivity \times Position interaction, $F_{1,212} = 13.02$, $p < .001$, $\eta_g^2 = .01$, indicating that the information about the additional machines affected appropriateness ratings in the sensitive condition more than in the insensitive condition. More importantly, this interaction was again qualified by a significant three-way Sensitivity \times Position \times Relationship interaction, $F_{1,212} = 12.23$, $p < .001$, $\eta_g^2 = .003$, indicating that the selective decrease of ratings from pre to

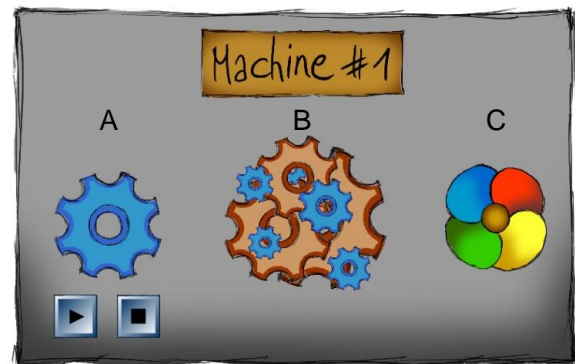


Figure 3: Screenshot of the animation used in the first learning phase of Experiment 2. Devices A and B were reversed in the Simple Mechanism condition. Letters A, B, and C were not shown to participants.

post in the sensitive condition was more pronounced for the $A \rightarrow C$ relationship than for the $B \rightarrow C$ relationship, just as in Experiment 1. The assignment of the single wheel and the complex system of wheels to distal cause (A) vs. proximal cause (B) did not affect the results, nor did this factor interact with any of the other variables in the design. The data shown in Figure 4 is therefore collapsed across this factor.

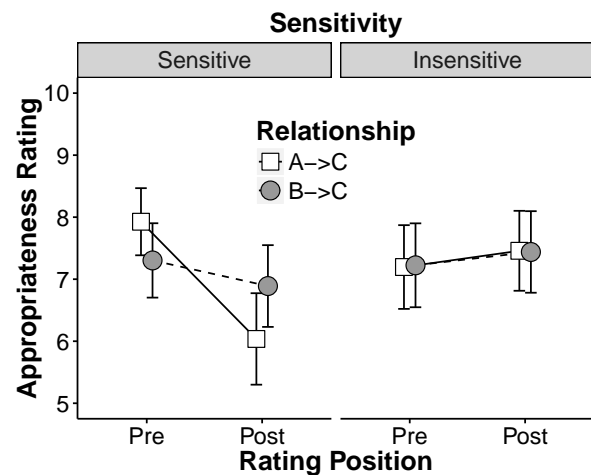


Figure 4: Group means (error bars = 95% CI) of appropriateness ratings in Experiment 2.

Subjects’ contingency estimates were analyzed analogous to Experiment 1 and yielded analogous effects. The $\Delta P(A \rightarrow C)_{\text{Exemplar}}$ estimates in the sensitive condition were again as high as the $\Delta P(B \rightarrow C)_{\text{Exemplar}}$ estimates. As before, this rules out that the selective drop in $A \rightarrow C$ appropriateness ratings results from decreased $A \rightarrow C$ dependency estimates within the focal entity.

Discussion

In this experiment, we have shown that the finding that sensitive mechanisms lead to devaluation of distal causes generalizes to the domain of artifacts. If setting a physical device in motion (A) leads to movement of a second device (B), which in turn sets in motion a third device (C), both A and B are seen as equally crucial for the movement of C. However, if it is later learned that this dependency does not generalize to other exemplars of the same kind of machine, people revise their interpretation of the chain even within the first exemplar. They now judge the distal cause to be less crucial for the occurrence of the effect than before, and as less crucial than the more proximal cause that mediates the relationship. The effects were even cleaner than in Experiment 1, despite the fact that we made causes A and B more similar to each other and even counterbalanced their contents. This supports our hypothesis that high sensitivity of mechanisms *per se* leads to an interpretation of the mechanism as alternative explanation rather than as a mediator of the original relationship.

General Discussion

In two experiments, we have demonstrated that sensitive mechanisms tend to be seen as alternative explanations of their effects rather than as mediators of the causal influence of a distal cause. When people learn about a new indirect causal relationship in a single context, their default intuition seems to be that the distal cause is a crucial contributor to the terminal effect. However, if they afterwards realize that the mechanism generating the dependency between distal cause and terminal effect breaks down in most other similar contexts, they revise this intuition and devalue the causal contribution of the distal cause. The information that the mediating mechanism requires highly specific, uncommon boundary conditions makes clear that the observed $A \rightarrow C$ dependency is highly sensitive (Woodward, 2006). Sensitive causes, in turn, tend to be regarded as somewhat deficient, presumably because they fail to support future predictions and interventions in similar cases (Lombrozo, 2010; Lombrozo & Carey, 2006). As Garfinkel (1981) put it, if we want to explain the occurrence of a particular outcome, our real object of explanation is never just the occurrence of *that particular* outcome. Instead, we search for stable causes that explain the occurrence of a whole equivalence class of inessentially different outcomes. If we find out that a mechanism relates a cause to a to-be-explained effect in only a small subset of cases within the relevant equivalence class (e.g., it only works reliably in a small number of exemplars of an otherwise apparently homogeneous kind), the cause does not provide a stable explanation for this kind of effect. The discovered mechanism then turns into an alternative explanation for the outcome that screens off the influence of the distal cause from the explanandum. This explanation captures not only the present data, but also previous findings in which sensitivity of the mechanism was not directly manipulated in statistical terms, but rather implied by qualitative characteristics of the described mecha-

nism (e.g., moral abnormality; see Nagel & Stephan, 2015).

The fact that sensitive mechanisms lead to a decrease in explanatory relevance of the distal cause A (rather than to an increase in relevance of the proximal cause B) suggests that the following psychological process might underlie the observed phenomenon. When sensitive mechanisms are observed, it becomes necessary to assume the influence of an additional, latent variable that interacts with the distal cause A to produce proximal cause B in a few but not all contexts (e.g., an abnormal preference structure of the teacher, or a genetic abnormality in Fish #1) in order to capture the structure of the complete situation. The apparent necessity of this additional variable for producing B (and, hence, C) makes it obvious that A is not sufficient in producing B (and, hence, C) even within the focal entity. Sufficiency, in turn, has been shown to be closely linked to explanatory relevance (e.g. Hilton, McClure, & Sutton, 2009; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). Future studies might aim to test this hypothesis more specifically.

Acknowledgments

We thank York Hagmayer, Jana Samland, and Michael Waldmann for helpful discussions.

References

- Garfinkel, A. (1981). *Forms of explanation: Rethinking the questions in social theory*. New Haven: Yale University Press.
- Hilton, D. J., McClure, J., & Sutton, R. M. (2009). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, 40, 383–400.
- Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, 79, 942–951.
- Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196–209.
- Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61, 303–332.
- Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99, 167–204.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 1–25.
- Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1691–1697). Austin, TX: Cognitive Science Society.
- Pearl, J. (2000). *Causation: Models, reasoning and inference*. Cambridge, MA: Cambridge University Press.
- Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, 115, 1–50.

Appendix J

Stephan, Willemsen, and Gerstenberg (2017)

Stephan, S., Willemsen, P. & Gerstenberg, T. (2017). Marbles in inaction: Counterfactual simulation and causation by omission. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1132-1137). Austin, TX: Cognitive Science Society.

Marbles in Inaction: Counterfactual Simulation and Causation by Omission

Simon Stephan¹ (sstephal@gwdg.de) Pascale Willemsen² (pascale.willemsen@rub.de)

Tobias Gerstenberg³ (tger@mit.edu)

¹Department of Psychology, Georg-August-University Göttingen and Leibniz ScienceCampus Primate Cognition

²Institute for Philosophy II, Ruhr-University Bochum

³Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

Consider the following causal explanation: The ball went through the goal because the defender didn't block it. There are at least two problems with citing omissions as causal explanations. First, how do we choose the relevant candidate omission (e.g. why the defender and not the goalkeeper). Second, how do we determine what would have happened in the relevant counterfactual situation (i.e. maybe the shot would still have gone through the goal even if it had been blocked). In this paper, we extend the counterfactual simulation model (CSM) of causal judgment (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014) to handle the second problem. In two experiments, we show how people's causal model of the situation affects their causal judgments via influencing what counterfactuals they consider. Omissions are considered causes to the extent that the outcome in the relevant counterfactual situation would have been different from what it actually was.

Keywords: causality; counterfactuals; causation by omission; causal attribution; mental simulation.

Introduction

Billy is on his way home. He is driving on a lonely country road, when he notices a damaged car next to the road. The car seems to have collided with a tree, and the driver appears unconscious. Billy decides not to stop and keeps driving. A few days later, Billy reads in the newspaper that the driver died because he had not received any medical attention.

Many people would concur that Billy's not having stopped was causally relevant for the driver's death. However, there are two fundamental problems with citing omissions (i.e., events that did not happen) as causes. First, there is the *problem of causal selection*. Why cite Billy's not stopping as causally relevant for the driver's death? Why not cite the Queen of England? Second, there is the *problem of underspecification*. Assuming that Billy would have stopped to check on the driver, what would he have done? Would Billy's acting have prevented the driver's death, or would she have died anyway?

In this paper, we show how the *counterfactual simulation model* (CSM) of causal judgment developed in Gerstenberg, Goodman, Lagnado, and Tenenbaum (2012) (see also Gerstenberg et al., 2014; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015) provides a natural solution to the underspecification problem. The CSM predicts that an omission is a cause when the positive event that is chosen as its replacement would have changed the outcome of interest. More specifically, we show how people's causal model of a situation guides their selection of the relevant counterfactual which subsequently determines their judgment about whether the omission made a difference to the outcome.

The paper is organized as follows: We first describe the causal selection and the underspecification problem in more detail. We then propose an extension to the CSM as a solution to the underspecification problem. Thereafter, we present and discuss the results of two experiments which test the CSM.

The Causal Selection Problem

Many philosophers argue that counterfactual approaches to causation are too inclusive when it comes to omissions (e.g. McGrath, 2005). If Billy had stopped and checked on the unconscious driver, the driver would not have died. Consequently, the driver died because Billy did not stop. However, following this logic, the same counterfactual seems to be true for the Queen of England. If the Queen of England had stopped, the driver would not have died either. However, intuitively, it is Billy's omission that was causally relevant, and not the Queen's. The problem of causal selection has been intensively discussed in both philosophy and empirical studies (e.g. Hesslow, 1988). Interestingly, while the causal selection problem presents a challenge to certain philosophical theories of causation, laypeople do not have any difficulty in selecting the cause of the driver's death. Based on evidence from research on causal cognition, it has been suggested that the concept of causation is not a purely descriptive one, but that it depends on reasoners' expectations (Willemsen, 2016). While we would have expected Billy to stop and help, we didn't entertain any such expectation for the Queen.

The Underspecification Problem

When it comes to omissive causation a fundamental problem is how to define the relevant counterfactual contrast (cf. Schaffer, 2005). For positive events ("something happened"), the counterfactual contrast ("it didn't happen") is often well-defined. However, replacing a negative event with a positive event seems more problematic because there are a infinitely many ways in which events can come about. If Billy actually helped the driver, it seems to be pretty clear what would have happened if he had not helped (he would just have continued to drive on). However, if Billy did not help, it is unclear what would have happened if he had helped (would he have helped in a competent manner to prevent the driver's death, or would he have been too nervous and screwed things up?).

While the causal selection problem has received much attention in the literature (e.g., Henne, Pinillos, & De Brigard, 2015; Livengood & Machery, 2007), the underspecification problem has not. One exception is the account by Wolff, Barbey, and Hausknecht (2010) that addresses both problems. The general idea proposed by Wolff et al. (2010) is that cau-

sation by omission is linked to the removal of an actual (or anticipated) force that previously prevented a certain outcome from occurring. One problem of this account, however, is that it appears too restrictive in that it cannot account for cases in which no (apparent) force is removed. Imagine, for instance, sentences like “The lack of rain caused the drought in Somalia”. Here, it would be a stretch to think of a the lack of rain as the removal of a force.

The extension of the CSM that we propose in this paper provides a different solution to the *underspecification problem*. Previous research has suggested that the extent to which a certain counterfactual is relevant is a function of both how likely we are to consider it, and how likely it would have changed the outcome of interest (Petrocelli, Percy, Sherman, & Tormala, 2011). However, while this research has shown that these counterfactual probabilities affect people’s causal judgments, it doesn’t explain how we come up with the relevant probabilities in the first place. Here, we will show how the CSM provides a natural solution to determine whether an omission made a difference to the outcome.

Counterfactual Simulation and Omission

The CSM predicts that people make causal judgments by comparing what actually happened with the outcome of a counterfactual simulation. So far, the model has been applied to capturing participants’ judgments about events that actually happened (Gerstenberg et al., 2012, 2014, 2015). Consider the situation shown in Figure 1b (bottom) illustrated as the *ideal path*. Here, A collides with B and B subsequently goes through the gate. The CSM says that ball A’s colliding with ball B caused ball B to go through the gate in this case, because it is obvious that ball B would have missed the gate but for the collision with A. More generally, the CSM predicts that causal judgments are a function of the reasoner’s *subjective degree of belief* that the candidate cause made a difference to the outcome. More formally, we can express the degree of belief that x caused y as

$$P(x \triangleright y) = P(y' \neq y | \mathcal{S}, \text{do}(x')), \quad (1)$$

in which x denotes the event of ball A hitting ball B, and the outcome y captures the event of ball B going through the gate. We first condition on what actually happened \mathcal{S} (i.e., the motion paths of each ball, the position of the walls, etc.). We then intervene to set the candidate cause event x to be different from what it was in the actual situation, $\text{do}(x')$. Finally, we evaluate the probability that the outcome in this counterfactual situation y' would have been different from the outcome y that actually happened. The results of several experiments (cf. Gerstenberg et al., 2012, 2014, 2015) have revealed that there exists a tight relationship between the counterfactual judgments of one group of participants (about what would have happened if the candidate cause had been absent), and the causal judgments of another group of participants.

To model causal judgments about positive events, the CSM considers counterfactuals in which the positive event (ball A colliding with B) is simply removed from the scene (indicated

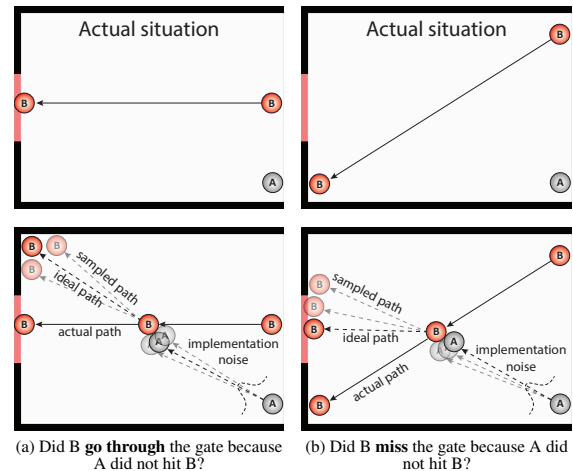


Figure 1: Illustration of what actually happened (top) and the counterfactual simulation model (bottom). The diagrams illustrate the actual path that ball B took, as well as an ideal path for (a) A preventing B from going through the gate, or (b) A causing B to go through the gate. The sampled paths show example simulations that result from applying implementation noise to the ideal path. *Note:* In (a), A would have prevented B from going through the gate for both sampled paths. In (b), A would have caused B go through the gate in one sample but not so in the other in which B would still have missed even though A hit B.

by $\text{do}(x')$ in Equation 1). Things become more intricate, however, when we want to model omissions as causes. As discussed above, it is often straightforward to replace an event with a non-event (e.g., a collision with no collision), but it is less clear how to replace a non-event with an event. Consider the situation shown in Figure 1a. Did ball B go through the gate because ball A *did not* hit it? The problem is that there are infinitely many ways for ball A to collide with ball B. Which of these events are we to consider? The collision event is severely underspecified. We will now show how the CSM can be extended to yield predictions about omissions as causes, and thereby provide a solution to the underspecification problem.

Modeling Omissions

We assume that people solve the underspecification problem by sampling counterfactual possibilities based on their intuitive understanding of the situation (cf. Kahneman & Tversky, 1982). The extent to which the omission is viewed as a cause of the outcome is assumed to be a function of the proportion of samples in which the outcome would have been different from what actually happened, assuming that the type of counterfactual event of interest was realized. Let us illustrate how the model works by example of the situation depicted in Figure 1a. In the actual situation, ball A did not move and ball B went right through the middle of the gate. We want to determine to what extent A’s not hitting ball B was a cause of B’s going through the gate. To do so, we simulate what would have happened if ball A had collided with B. More specifically, we need to determine the time t at which A would have started to move, the direction d in which ball

A would have moved, and the velocity v . Once we have determined these quantities, we can simulate what would have happened. For many combinations of values for t , d , and v ball A would not have collided with ball B. We can discard all such situations since we are interested in evaluating what would have happened if ball A had *hit* ball B. For each situation in which the two balls collide, we record what the outcome would have been – would B have missed the gate, or would it still have gone through the gate? We can now obtain the probability that ball A’s not hitting ball B was a cause of ball B’s going through the gate (cf. Equation 1) by looking at the proportion of samples in which B would have missed the gate instead of going through.

But how do we determine what values to take for t , d , and v which jointly determine what counterfactual situation we consider? We predict that prior expectations guide the counterfactuals we consider. In Experiment 1 below, we contrast situations in which participants don’t have any expectations about what normally happens, with situations in which participants have statistical, or social expectations. We will now discuss how the model incorporates these expectations.

Expectations Shape Counterfactual Simulations

No Expectations Let us first assume a situation in which an observer does not have any strong expectations concerning how the balls typically move in the given context. When asked whether A’s not hitting B caused B to go through the gate, we have to generate situations in which A would have hit B. This already considerably constrains what kinds of situations we consider. For example, it would be futile to consider situations in which A only starts moving after B already went through the gate, or in which A moved toward the right.

We generated counterfactual samples in the following way: We first discretized the space for the time at which A starts moving t , the direction in which it moves d , and its velocity v . For t , we considered all values from 0 to $t_{outcome}$ where 0 corresponds to the time at which B starts moving and $t_{outcome}$ to the time at which ball B went through the gate (or hit the wall). For d , we considered the full range from A going straight to the left to going straight up. For v , we considered a reasonable range from A moving slowly to A moving fast. For each generated world, we noted whether A and B collided, and whether B went through the gate or missed the gate. We then discarded all situations in which the two balls did not collide, and recorded the proportion of situations in which B would have gone through the gate if the balls had collided.

The model makes the following predictions: For the situation in which B is on a path toward the gate (Figure 1a), there is a good chance that B would have missed the gate if ball A had hit it. The model predictions are shown in Figure 2. As can be seen in the left panel, the CSM concludes that the probability that B would have missed the gate had A hit it is just as high as the probability that B would have passed the gate. By contrast, when B is on a path away from the gate (“missed” in Figure 2, cf. Figure 1b top right) there is only a

relatively small chance that ball B would have gone through the gate if ball A had hit it. Thus, the CSM predicts that people will be more likely to agree that ball B *went through* the gate because ball A did not hit it than they will be to agree that ball B *missed* the gate because ball A did not hit it.

Social Expectations When nothing particular is known about how A and B typically move, the space of counterfactuals from which the CSM samples is relatively wide. It seems plausible, however, that what counterfactual possibilities are considered will be affected by different forms of prior expectations. Imagine, for example, that you learn that two players play a marble game. Player B wants to get her marble into the goal, while Player A wants to make sure that this does not happen. On a particular trial, Player A did not pay attention and forgot to flick his marble. Did Player B’s marble go through the gate because Player A’s marble did not hit it? When knowing that it is a player’s job to prevent a marble from going through the gate, people may expect that this player would not have just flicked her marble randomly. Instead, she can be expected to try her best to make sure that the other marble does not go through the gate. Similarly, consider a situation in which Player A also wants that Player B’s marble goes through the gate. In that case, it seems likely that Player A will try to flick his marble so that it makes sure that B’s marble will go through the gate.

Figure 1 illustrates how the CSM incorporates how prior expectations constrain the space of counterfactual situations. We assume that the player would first determine a time t at which to flick her marble. For any given point t , the player then determines an optimal d and v conditional on the player’s goals. For a player who wants to prevent ball B from going through the gate, the player’s goal is to maximize the distance between B’s position and the middle of the gate. For a player who wants to cause B to go through the gate, the player’s goal is to minimize the distance between B’s position and the middle of the gate (i.e., she wants B to go right through the middle of the gate). For simplicity, we assume that players can plan their action optimally, but that they have some implementation noise. The CSM models this implementation noise by introducing a small perturbation to the ideal path on which A moves. As is illustrated in Figure 1, the CSM incorporates implementation noise by slightly perturbing the “ideal path” vector.

Figure 1 shows the actual path that ball B took, the ideal paths that player A “wanted” the marbles to take, and two examples for paths that ball B actually took after subjecting A’s ideal plan to some implementation noise. Notice that the implementation noise has a larger effect in Figure 1b where it leads to a situation in which ball B would have missed the gate even though ball A hit it. In contrast, in Figure 1a the implementation noise has less of an effect. Here, ball B would reliably miss the gate even if we apply some implementation noise to player A’s intended plan. Accordingly, the CSM predicts that it is more likely that A’s hitting B would have resulted in B missing the gate (when B actually went through,

Figure 1a) than it would have resulted in B going through the gate (when B actually missed, Figure 1b). Since the sample of considered situations is biased toward optimal actions, the CSM predicts that judgments will overall be higher than when an observer does not have any prior expectations. The predictions for this situation are shown in the middle panel in Figure 2.

Statistical Expectations Now imagine that instead of learning anything about agents playing a game you get to see a few situations first that shape your expectations about what tends to happen. We incorporate such “statistical” expectations into the model in the same way in which we handled social expectations. However, we allow for the implementation noise to be different between these situations. Specifically, the size of the implementation noise parameter will depend on the kind of evidence that participants have seen. For example, if one has witnessed a series of trials in which A always hit B in such a way that B went straight through the gate, this would suggest a smaller implementation noise compared to one that is suggested by trials in which A hit B in such a way that B went through the gate in, for example, merely two third of the cases. The predictions for this situation are shown in the right panel in Figure 2.

Experiment 1

Experiment 1 tests whether the CSM accurately predicts people’s causal judgments for omissions in dynamic physical scenes. We look at causal judgments about situations in which ball A failed to hit ball B, and ball B either went through or missed the gate (see Figure 1). In line with the CSM, we predict that the degree to which people judge ball A’s not hitting ball B as causally relevant to the outcome would be tightly coupled with the results of a mental simulation about what would have happened if a collision had occurred. Furthermore, we test the hypothesis that different types of expectations (social or statistical) influence people’s causal judgments by affecting what counterfactual situations people consider.

Methods

Participants and Materials 476 participants (239 female, $M_{Age} = 33.83$ years, $SD_{Age} = 12.03$ years) were recruited via Prolific Academic (www.prolific.ac) and participated in this experiment for a monetary compensation of £0.25. The clips were created in Adobe Flash CS5 using the physics engine Box2D.

Design and Procedure All factors were manipulated between subjects. We manipulated what actually happened (*actual outcome*: missed vs. went through), and the expectations of participants about what will happen (*expectation*: no expectations, statistical expectation, social expectation). Finally, we varied whether participants answered a causal question, or a (counterfactual) probability question (*question*: causation vs. probability).

In the “no expectations” condition, subjects simply read that they will see an animation in which a stage with solid

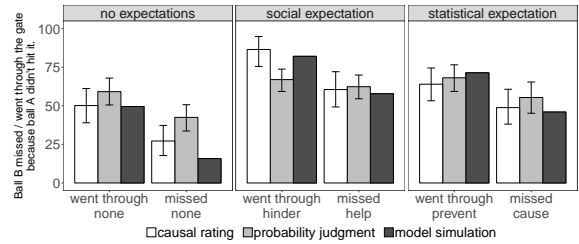


Figure 2: **Experiment 1.** Mean causal and probability judgments together with the predictions of the CSM. *Note:* Error bars indicate 95% bootstrapped CIs.

walls, two balls A and B, and a gate will be displayed. All subjects were shown a graphical illustration of the stimuli. Participants in the “statistical expectation” condition were presented four primer clips in which ball B actually collided with A. One group of subjects saw that the collision always *caused* B to go through the gate, while the other half always saw that A *prevented* B from going through the gate (see Figure 1). In the “social expectation” condition, subjects were instructed that the video clip (which was the same as in the “no expectations”) shows what happened during a game of marbles played by two agents, Andy and Ben. We manipulated whether subjects believed that Andy wants to *help* Ben to flip his marble through the gate or whether he wants to *hinder* Ben from doing so.

Participants in the “causation” condition indicated how much they agreed with the claim that B *missed* the gate because A did not hit it, or that B *went through* the gate because A did not hit it, depending on the outcome. Participants in the “probability” condition gave a corresponding probability judgment: they indicated what they believed the chances were that B would have gone through / missed the gate if ball A had hit ball B. Participants indicated their ratings on a sliding scale.

Which outcome participants saw depended on the expectation condition: In the “social expectation” condition, participants who expected the agent to *help* saw that B actually missed the gate, and participants who expected the agent to *hinder* saw that B went through the gate. In the “statistical expectations” condition, participants who had seen the *causation* clips saw that B missed the gate, whereas those who had seen the *prevention* clips saw that B went through the gate.

Results and Discussion

Figure 2 shows participants’ mean causal ratings (white bars), probability ratings (gray bars), as well as the predictions of the CSM (black bars). The CSM correctly predicts a difference in agreement ratings for both the causal and probability condition as a function of the outcome (went through vs. missed). A global 2 (question) \times 6 (combination of expectation and outcome) factorial ANOVA shows a main effect of outcome, $F(5,464) = 14.51$, $p < .001$, $\eta_G^2 = .61$ but no main effect of question, $F(1,464) < 1$. The interaction between question and expectation was significant,

$F(5, 464) = 2.74, p < .05$ but the effect is small, $\eta_G^2 = .03$.

Importantly, participants saw A's not hitting ball B as more causal when B went through the gate compared to when it missed. This pattern was predicted by the CSM and indicates that participants' counterfactual simulations and their causal inferences were sensitive to the constraints imposed by the virtual physical environment. Because the displayed gate was relatively small, the probability that a collision would change the outcome is higher if B actually went through, than when it missed. Planned contrasts confirmed that the observed differences between "went through" and "missed" were significant in all conditions, with $t(464) = 3.21, p < .01, r = .15$ in the "no expectations" condition, $t(464) = 2.13, p < .05, r = .10$ in the "statistical expectation" condition, and $t(464) = 3.53, p < .001, r = .16$ in the "social expectation" condition.

Besides the asymmetry between "went through" and "missed", we also expected to see higher causality ratings in the "statistical" and "social expectation" conditions than in the "no expectation" condition. This difference was predicted because we incorporated an ideal path in these situation that was then perturbed by imposing some implementation noise. As Figure 2 shows, we did indeed observe this pattern. A planned contrast confirmed that this difference was significant, $t(464) = 5.98, p < .001, r = .27$.

Concerning the probability ratings, planned contrasts showed that the difference between "went through" and "missed" was significant in the "no expectations condition", $t(464) = 2.33, p < .05, r = .11$, and the "statistical expectation" condition, $t(464) = 1.73, p < .05, r = .08$, but not in the "social expectation" condition, $t(464) < 1$. Concerning the predicted difference between the "no expectations" condition and the other two expectation conditions, Figure 2 shows that we obtained a similar pattern as for the causality judgments. In line with our expectations, the probability ratings for the "statistical expectation" and the "social expectation" condition were higher than the ratings for the "no expectation" condition, $t(464) = 2.82, p < .01$, though this effect was smaller than the effect for the causality judgments, $r = .13$.

The results of Experiment 1 show that participants' causal judgments are qualitatively well accounted for by the CSM. The CSM also does a good job in accounting for the pattern quantitatively, as evidenced by a high correlation between model predictions and counterfactual probability judgments ($r = .97, RMSE = 14.00$), as well as between model predictions and causal judgments ($r = .97, RMSE = 6.06$). The fact that the model accounts slightly less well for the counterfactual probability judgments is mainly due to the relatively large difference between model predictions and probability judgments in the "no expectations" condition.

A key finding in Experiment 1 is the asymmetry in participants' causal judgments as a function of whether ball B went through or missed the gate. The CSM predicts this pattern because it is more likely that A's hitting B would prevent B from going through the gate (cf. Figure 1a) than that it would cause

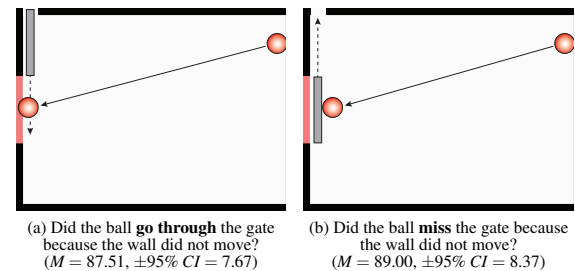


Figure 3: Illustration of the materials used in Experiment 2. Solid arrows indicate the actual path of the ball; dashed arrows show the hypothetical path of the wall. Graph (a) shows the "went through" and (b) the "missed" condition. The results for both conditions are included in brackets.

B to go through (cf. Figure 1b). One possibility, however, that Experiment 1 cannot rule out is that people are in general more likely to regard omissions as causes when the relevant counterfactual involves preventing compared to causing. In Experiment 2, we investigate whether there is such a general asymmetry between omissive causation and prevention.

Experiment 2

The goal of Experiment 2 was to rule out that the observed difference between "went through" and "missed" in Experiment 1 came about because people generally treat omissive causation and omissive prevention differently. The CSM only predicts an asymmetry between two situations when the positive event of interest was more likely to make a difference in one situation compared to the other. Hence, our strategy in Experiment 2 was to hold this probability constant. To achieve this goal, we simply replaced ball A with a wall that had exactly the size of the gate. To model "missed" and "went through", we varied whether the wall blocked the gate or not, while a displayed ball always headed toward the gate (see Figure 3). Participants rated how much they agree that "the ball" missed the gate (or went through the gate) because the wall did not move. There is no ambiguity about the relevant counterfactual in this case – it is clear that the outcome would have been different, had the wall moved. Accordingly, the CSM predicts that participants' judgments should be high for both cases, no matter whether the ball went through the gate or missed the gate because of the omission.

Method

Participants 65 participants (40 female, $M_{age} = 32.86, SD_{age} = 12.84$) who were again recruited via Prolific Academic completed this online experiment and received a monetary compensation of £ 0.25.

Design, Materials, and Procedure The final outcome, that is, whether the ball went through or missed the gate (see Figure 3) was manipulated between subjects. The instructions were similar those used in the "no expectations" condition in Experiment 1. Further, participants were presented an illustration showing the materials in which it was made clear that the wall can only be in two different positions, either right

in front of the gate or in the upper left corner of the stage (see Figure 3). Having read the instructions, participants were shown the respective video clip and provided the causal rating after the clip was finished.

Results and Discussion

As expected, participants gave very high causal ratings for “went through” ($M = 87.51$, $SD = 21.62$) and “missed” ($M = 89.00$, $SD = 23.21$). As predicted by the CSM, the ratings were not different from each other, $t(63) < 1$. The probability that the outcome would have been different in the relevant counterfactual, is close to maximal in both conditions.

The results of Experiment 2 are in line with the CSM. Further, the fact that the causality ratings were both very high and not different from each other rules out the potential alternative explanation that people might generally treat omissive causation and omissive prevention differently.

General Discussion

We developed an extension of the *Counterfactual Simulation Model* to account for causation by omission. Based on previous research by Gerstenberg et al. (2014), we reasoned that people’s causal judgments are closely linked to their subjective degree of belief that the outcome would have been different had the candidate cause been replaced. We argued that this replacement by a counterfactual contrast is particularly difficult in cases of omissions. The counterfactual contrast to “did not hit” is clearly “had hit”, but it remains unclear what would have happened if “hitting” had taken place.

In two experiments we shed light on how to tackle the underspecification problem. We predicted that prior expectations would constrain what counterfactual contrasts people consider relevant to the scenario. Experiment 1 revealed an asymmetry: A’s not hitting B was judged less causal when B missed the gate compared to when B went through the gate. This is what the CSM predicts, and the results thus lend additional support to the hypothesis that causal judgments are grounded in counterfactually simulated probabilities. Adding expectations increased both people’s causal judgments as well as their subjective degree of belief that a counterfactual collision would have changed the outcome. This effect was particularly strong for social expectations, which the CSM explains by assuming that knowledge about intentions of agents limits the range of counterfactuals that are considered. Our results thus add to previous research indicating that intentional actions signal higher causal stability compared to unintentional ones (Lombrozo, 2010), and that causal stability is indeed a relevant dimension that affects causal reasoning (Nagel & Stephan, 2016).

It might be objected that the asymmetry in causal attribution for “went through” and “missed” in Experiment 1 is not due to a difference in what would have happened in the relevant counterfactual simulations, but rather due to an inherent asymmetry between omissions that prevent and omissions that cause. Experiment 2 addressed this possible con-

found by looking at situations in which the relevant counterfactual event was clear (a wall that could only move in one direction), as well as what would have happened in case that event had happened. Just as predicted, we found that causal ratings were equally high irrespective of whether the ball “went through” and “missed” in this case. Instead of a general asymmetry between prevention and causation, participants judge omissions to be causal the more certain they are that the omission made a difference to the outcome.

As our introductory example demonstrates, omissions are particularly relevant in human interaction, especially so in morally or legally charged situations when we had clear expectations about what a person should have done. In this paper, we have shown how the CSM accounts for people’s causal judgments of omissions in situations in a physical domain in which the relevant counterfactuals are relatively well constrained. However, we believe that the CSM has the potential to capture causal judgments about omissions of social agents as well. For example, the extent to which we blame someone for not having helped depends on how easy it would have been for the agent to help (cf. Jara-Ettinger, Tenenbaum, & Schulz, 2015). In future research, we will explore the CSM in a richer social setup.





Acknowledgments This project was supported by the Leibniz Association through funding for the Leibniz ScienceCampus Primate Cognition. TG was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216.

References

- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 523–528). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.
- Henne, P., Pinillos, Á., & De Brigard, F. (2015). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, 1–14.
- Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). Brighton, UK: Harvester Press.
- Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers’ inferences about costs and culpability. *Psychological Science*, 26(5), 633–640. Retrieved from <http://dx.doi.org/10.1177/0956797615572806> doi: 10.1177/0956797615572806
- Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.
- Livengood, J., & Machery, E. (2007). The folk probably don’t think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy*, 31(1), 107–127.
- Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4), 303–332.
- McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, 123(1), 125–148.
- Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In A. Papafragou, D. Grodner, D. Miram, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 806–812). Austin, TX: Cognitive Science Society.
- Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology*, 100(1), 30–46.
- Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, 114(3), 327–358.
- Willemsen, P. (2016). Omissions and expectations: A new approach to the things we failed to do. *Synthese*. Advance online publication. doi: 10.1007/s11229-016-1284-9
- Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, 139(2), 191–221.

Simon Stephan

Curriculum Vitae

 Department of Psychology, University of Göttingen, Gosslerstr. 14, D-37073 Göttingen
 +49 551 39 33762
 sstephal@uni-goettingen.de
 www.psych.uni-goettingen.de/de/cognition/team/stephan

EDUCATION

- SINCE 2015 **PhD Student in the Program: “Behavior and Cognition”**
 PSYCHOLOGY
 Georg-August-University Göttingen,
 Thesis Supervisor: Michael R. Waldmann
- SINCE 2014 **Research Assistant at the Cognitive and Decision Sciences Lab**
 DEPARTMENT OF PSYCHOLOGY
 Georg-August-University Göttingen
- 2012 – 2014 **Master of Science (with distinction)**
 PSYCHOLOGY
 Georg-August-University Göttingen
- 2009 – 2012 **Bachelor of Science (with distinction)**
 PSYCHOLOGY
 Georg-August-University Göttingen
- 2002 – 2009 **Abitur (eqvl. A-levels)**
 Grotefeld-Gymnasium Münden, Hann. Münden

AWARDS & SCHOLARSHIPS/GRANTS

- OCTOBER 2017 **DFG Grant**
 VALUE: 212,768.00€
 Project: “Answering causal queries about singular cases” (The official holder of this grant is Michael R. Waldmann)
- JULY 2017 **Computational Modeling Prize in High-level Cognition**
 Sponsored by the Cognitive Science Society for the best full paper submissions that involve computational cognitive modeling.
- 2016 – 2018 **Leibniz-ScienceCampus Grant**
 VALUE: 9,552.80€
 Project: “The relationship between causal and moral judgments”
- 2015 – 2017 **Leibniz-ScienceCampus Grant**
 VALUE: 7,272.40€
 Project: “The role of intentions in children’s and adult’s causal ascriptions”
- 2011 – 2012 **e-fellows.net Scholarship**

PUBLICATIONS

- Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2018). Assessing singular causation: The role of causal latencies. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings*

- of the 40th Annual Conference of the Cognitive Science Society* (pp. 1080-1085). Austin, TX: Cognitive Science Society.
2. Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, 10, 242–257.
 3. Stephan, S., & Waldmann, M. R. (2017). Preemption in Singular Causation Judgments: A Computational Model. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1126–1131). Austin, TX: Cognitive Science Society.
 4. Stephan, S., Willemsen, P., & Gerstenberg, T. (2017). Marbles in inaction: Counterfactual simulation and causation by omission. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. (pp. 1132-1137). Austin, TX: Cognitive Science Society.
 5. Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 806-812). Austin, TX: Cognitive Science Society.
 6. Stephan, S., & Waldmann, M. R. (2016). Answering causal queries about singular cases. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2795-2801). Austin, TX: Cognitive Science Society.
 7. Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1691-1696). Austin, TX: Cognitive Science Society.

Submitted articles

8. Stephan, S., & Waldmann, M. R. (submitted). The role of causal power and temporal information in singular causation judgments.

Articles in preparation

9. Stephan, S., Tentori, K., Pighin, S. & Waldmann, M. R. (in preparation). Interpolating causal mechanisms: The paradox of knowing more.

CONFERENCES

-
- JULY 2018 **Cognitive Science Conference (40th), Madison, USA**
Talk: “Assessing Singular Causation: The Role of Causal Latencies”
- MAY 2018 **International Meeting of the Psychonomic Society, Amsterdam, NL**
Poster: “Answering Singular Causation Queries: The Role of Temporal and Mechanistic Information”
- FEBRUARY 2018 **Annual Meeting (7th) of the DFG Priority Program “New Frameworks of Rationality”**
- AUGUST 2017 **European Society for Philosophy and Psychology (ESPP) Conference, Hertfordshire, UK**
Talk: “Answering causal queries about singular cases”
- JULY 2017 **Cognitive Science Conference (39th), London, UK**
Talk: “Preemption in singular causation judgments: A computational model”

- JULY 2017 **Cognitive Science Conference (39th), London, UK**
Talk: "Marbles in inaction: Counterfactual simulation and causation by omission"
- MARCH 2017 **Annual Meeting (6th) of the DFG Priority Program "New Frameworks of Rationality"**
Talk: "Answering causal queries about singular cases"
- AUGUST 2016 **Cognitive Science Conference (38th), Philadelphia, USA**
Talk: "Answering causal queries about singular cases"
- AUGUST 2016 **Cognitive Science Conference (38th), Philadelphia, USA**
Poster: "Explanations in causal chains: Selecting distal causes requires exportable mechanisms" (presented by Jonas Nagel)
- JULY 2015 **Cognitive Science Conference (37th), Pasadena, USA**
Poster: "Mediators or alternative explanations: Transitivity in human-mediated causal chains" (presented by Jonas Nagel)

TEACHING EXPERIENCE

Tutorials/ Seminars

- WINTER TERM 2014/2015/2016 **Tutorial in "Quantitative Methods I"**
Part of the first year undergraduate psychology statistics class
- SUMMER TERM 2015/2016/2017 **Tutorial in "Quantitative Methods II"**
Part of the first year undergraduate psychology statistics class

Supervision

- WINTER TERM 2018/19 **Supervision of a Bachelor Thesis**
on the interpolation of causal chains
- SUMMER TERM 2018 **Supervision of a Bachelor Thesis**
on the role of category levels in causal judgments
- SUMMER TERM 2018 **Supervision of a Bachelor Thesis**
on the influence of statistical norms on causal selection
- SUMMER TERM 2017 **Supervision of a Bachelor Thesis**
on preemption in singular causation judgments
- WINTER TERM 2016/17 **Supervision of a Bachelor Thesis**
on singular causation judgments
- WINTER TERM 2016/17 **Supervision of a Bachelor Thesis**
on causal reasoning about double prevention

REVIEWS

- COGNITIVE SCIENCE CONFERENCE PROCEEDINGS: 6
 PSYCH REVIEW (AS CO-REVIEWER): 1

SKILLS

- LANGUAGE German (native), English (fluent), French (basic)
- SOFTWARE R, Photoshop, Illustrator, Flash, Animate, \LaTeX , HTML5
- INTERESTS Football, Badminton, Politics, Philosophy, Literature, Music, Guitar, Blues Harp