AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2021, Vol. 150, No. 8, 1500–1527
https://doi.org/10.1037/xge0001016

Interpolating Causal Mechanisms: The Paradox of Knowing More

Simon Stephan[1], Katya Tentori[2], Stefania Pighin[2], and Michael R. Waldmann[1]
[1] Department of Psychology, University of Göttingen
[2] Center for Mind and Brain Sciences (CIMeC), University of Trento

Causal knowledge is not static; it is constantly modified based on new evidence. The present set of seven experiments explores 1 important case of causal belief revision that has been neglected in research so far: *causal interpolations*. A simple prototypic case of an interpolation is a situation in which we initially have knowledge about a causal relation or a positive covariation between 2 variables but later become interested in the mechanism linking these 2 variables. Our key finding is that the interpolation of mechanism variables tends to be misrepresented, which leads to the paradox of knowing more: The more people know about a mechanism, the weaker they tend to find the probabilistic relation between the 2 variables (i.e., weakening effect). Indeed, in all our experiments we found that, despite identical learning data about 2 variables, the probability linking the 2 variables was judged higher when follow-up research showed that the 2 variables were assumed to be directly causally linked (i.e., C→E) than when participants were instructed that the causal relation is in fact mediated by a variable representing a component of the mechanism (M; i.e., C→M→E). Our explanation of the weakening effect is that people often confuse discoveries of preexisting but unknown mechanisms with situations in which new variables are being added to a previously simpler causal model, thus violating causal stability assumptions in natural kind domains. The experiments test several implications of this hypothesis.

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

Research focusing on causal reasoning generally explores how people acquire and use knowledge about relations between causes and effects. Many theories share the assumption that the world can be categorized into a given set of variables that can be arranged in causal networks. 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). Less attention has been paid to the dynamic process of extending and deepening our knowledge (for notable exceptions, see Bramley et al., 2017; Oaksford & Chater, 2013, 2017; Taylor & Ahn, 2012). People do not only learn about causal relations between variables; they also acquire knowledge about the mechanisms mediating previously discovered covariations or 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 heart disease or that a new drug relieves headache 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 interpolation of causal networks.

The process of rerepresenting our causal knowledge into more elaborate causal models can be expressed in various ways. A popular theoretical framework 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 Figure 1 for an example; see Rottman, 2017; Rottman & Hastie, 2014; Waldmann, 2017; Waldmann & Hagmayer, 2013, for reviews). The building blocks of causal Bayes nets are direct causal relations between causes and effects, but these direct relations can be combined into indirect ones forming causal chains or more complex kinds of networks (see Figure 1).

In a given stage of our knowledge acquisition process, we may have constructed the representation of a specific causal network. However, even when we are confident that our model is adequate, we may still want to elaborate the model by adding new variables. It is important to note that the direct causal relations within a causal model are only direct relative to a specific set of variables. Indeed, causal models are *frame-relative* in the sense that it is always possible to turn a direct causal relation into an indirect one by *interpolating* new variables that mediate the previously directly linked ones (see Spohn, 2012). What is represented as a direct causal relation by some people may be represented as an indirect one by other people. For example, most people will represent the relation between the intake of aspirin and the relief of headache as a direct causal relation. However, some of us who are interested in

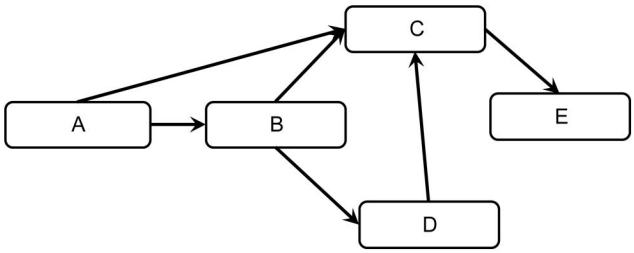
This article was published Online First February 1, 2021.

Simon Stephan https://orcid.org/0000-0002-6557-9637
Michael R. Waldmann https://orcid.org/0000-0002-8831-552X

This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (WA 621/24-1) and by the Leibniz Association through funding for the Leibniz ScienceCampus Primate Cognition. The senior author received a grant from the University of Trento for a stay as a visiting scholar. We thank Mike Oaksford, Andrea Passerini, and Bob Rehder for helpful comments. The materials and data of the reported studies can be accessed at https://osf.io/aqzps/.

Correspondence concerning this article should be addressed to Michael R. Waldmann, Department of Psychology, University of Göttingen, Gosslerstrasse 14, 37073 Göttingen, Germany. Email: michael.waldmann@bio.uni-goettingen.de

Figure 1
Example of a Causal Bayes Net Depicting the Causal Relations Between the Five Variables A, B, C, D, and E



biology and medicine may want to know how aspirin achieves this effect. The mechanism involves the inhibition of COX enzymes preventing the conversion of arachidonic acid to prostaglandin. Within a Bayes net framework, mechanisms are finer-grained, more elaborate representations of a sequence of causal dependencies (Woodward, 2011). There is no natural end point here (Cartwright, 1999). A biochemist may even go further and explore further submechanisms linking arachidonic acid with prostaglandin. Thus, the same covariational relation between aspirin and headache can be represented with different sets of variables and degrees of resolution. What is direct within one framework may be indirect within others.

One major goal of our project is to investigate how people reason with these different levels of description and how people revise their causal beliefs after being confronted with new information about mechanisms that was unknown to them before. Thus, we will study how people reason in situations of causal belief revision. In most experiments we will focus on causal chains, but later we will demonstrate that our findings can be generalized to more complex structures. Our experiments will demonstrate that people's causal inferences are biased in a way that contradicts a normative causal analysis.

Revising Beliefs About Causal Chains

There are various ways in which knowledge about a causal chain can be revised. First, it could turn out that the causal directions within a chain do not adequately represent the real causal dependencies. Some of the covariations may, for example, be attributable to a common cause rather than a direct causal link (see Taylor & Ahn, 2012). 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 initial strength estimates were distorted and are in fact stronger or weaker. Third, new relevant variables connected to the known network could be discovered, which leads to an *augmented network*. In the case of causal chains, adding variables may lead to a *lengthening* of the chain. Alternatively, mechanisms mediating between variables could later be discovered, which leads to the *interpolation* of variables. Of course, these possibilities of belief revision can be combined.

Interpolations, in particular, have rarely been studied. As an example for an interpolation, our knowledge acquisition process may start with observations that bolster our belief in a stable causal contingency between intake of aspirin and relief of headache. In this initial causal model representation, aspirin would play the role of a direct preventive cause of headache with the causal strength parameter reflecting the observed probabilities. Later we may discover how this relation is mediated. This would lead to a more elaborate representation turning the direct causal relation (aspirin → relief of headache) into an indirect one (e.g., aspirin → prostaglandin → relief of headache). In this example, other components of the mechanism (see above) are yet unknown to the reasoner so that only one mechanism variable is interpolated. The key question of the present study concerns how extensions of causal knowledge about chains affect our beliefs in the probabilistic relations between causal variables. In the following section, we will consider in depth the two cases of extending causal chains, *lengthening* and *interpolating*. Our experiments will then focus on interpolations, which have not yet been investigated in detail.

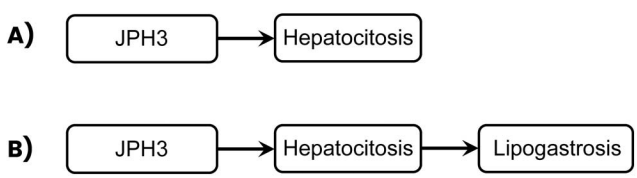
Lengthening of Causal Chains

Lengthening a given chain by adding new variables at the beginning or at the end of the chain typically leads to a weakening of probabilistic relations between the variables at the extremes proportional to their distance. A key assumption here is that the initial chain along with its causal strength parameters 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 Hepatocitosis (a fictitious disease; see Figure 2A). Because other known and unknown outside variables typically additionally affect causal relations, the causal relation between JPH3 and Hepatocitosis will most likely be probabilistic on the observational level. Then assume that it has been discovered that Hepatocitosis directly causes another disease, Lipogastrosis. This would imply a lengthening of the chain from two to three variables (see Figure 2A and 2B, respectively).

There are many ways to measure the strength of a covariation between variables, but a standard method is to use the contingency measure ΔP (see, e.g., Perales et al., 2017, for a review), which equals the difference between the conditional probabilities $P(E|C) - P(E|\neg C)$, with E representing the effect, C the target cause, and $\neg C$ the absence of the cause. ΔP can range from -1 (for a deterministic *preventive* relation) to $+1$ (for a deterministic *generative* relation).

With reference to Figure 2B, if we assume that the Markov condition holds (see Mayrhofer & Waldmann, 2015; Rehder & Waldmann, 2017), which in this case means that the new probabilistic relation between Hepatocitosis and Lipogastrosis is independent of whether JPH3 is present or absent, the following relation also holds:

Figure 2
Illustration of the Lengthening of an Initial Causal Chain (A) into a Causal Chain With an Additional Effect (B)



$$\Delta P_{JPH3-Lipo} = \Delta P_{JPH3-Hepa} \times \Delta P_{Hepa-Lipo} \quad (1)$$

Equation 1 states that the contingency between JPH3 and Lipogastrosis is equal to the product of the contingencies of the two directly linked variables (i.e., JPH3-Hepatocitosis, Hepatocitosis-Lipogastrosis), which leads to a *weakening* effect when the causal links are probabilistic. 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 because our focus will not be on lengthening but rather on 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). Presented with individual pairwise causal links that are later combined into a three-variable chain, participants 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 (Ahn & Dennis, 2000; Baetu & Baker, 2009; Jara et al., 2006). Interestingly, reasoning about chains even tends to express a transitivity bias when the presented data actually contradict the Markov condition and therefore are inconsistent with transitive chains (von Sydow et al., 2016, 2010). There seems to be a tendency to assume the validity of the Markov constraint even when it is violated, at least in these tasks.

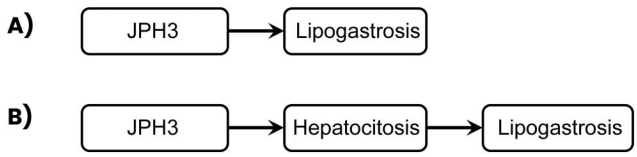
A study testing whether causal reasoning is consistent with the normative predictions of Bayes nets was conducted by Bes et al. (2012). The study investigated a number of causal models and showed judgment biases that violated normative assumptions underlying Bayes nets. However, some of the results that are relevant for our study seemed consistent with the multiplication rule. For example, in their Experiment 2, Bes et al. (2012) presented three causal variables that were sufficiently neutral so that they could be arranged in different causal models through verbal instructions. 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 the *direct predictive chain condition* A was the direct cause of B and B the direct cause of C (i.e., A→B→C), whereas in the *indirect predictive chain condition* B was indirectly caused by A via C (i.e., A→C→B). In a within-subject design, each participant was presented with the contrasted conditions, but the variables in the different conditions referred to different scenarios. In the test question measuring probabilistic intuitions, participants were requested to rate the probability of effect B given the presence of cause A (i.e., P(B|A)) on a scale ranging from 0% to 100%. The key finding was that, despite identical information about the contingencies between A and B, the 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 (Bes et al., 2012, Experiment 3) in which, prior to causal model instructions, trial-by-trial data were presented showing individual cases and in which the number of learning trials was manipulated. This effect was not sensitive to the length of the training phase. This study provides further evidence for participants' belief that causal chains are transitive although the data actually violated the Markov condition in the experiments.

An interesting question is whether the observed effect is normative. In general, there is no reason why in two different scenarios with different variables (as in the within-subject design of Bes et al., 2012) the direct relation between A and B (A→B), for example, should place quantitative constraints on the strength of a relation between the different variables C and E in a different indirect causal relation (C→D→E). The covariation between A and B can be larger or smaller than the covariation between C and E, or their sizes can be equal (as in Bes et al., 2012). If, for example, A and B correspond to the intake of aspirin and relieve from headache, and C and E to flipping a switch and the switching on of a lightbulb via an electric current (D), the probabilistic dependency between the indirectly linked variables, C and E, is surely higher than between the directly linked variables, A and B, in our world. One reason why Bes et al. (2012) may have observed a stable weakening effect consistent with the multiplication rule may have been that, in all their conditions, three-variable chains were presented with the direct causal relation always being a component of an indirect one. If, for example, participants learned in different within-subjects conditions about the chains A→B→C and D→E→F, the request to estimate conditional probabilities for the direct relations asked about the A→B or D→E relation, whereas the request to estimate indirect relations asked about the probabilities linking A and C or D and F. Thus, the direct relations were subcomponents of the indirect ones, thus suggesting that generally indirect relations were obtained through lengthening of the direct ones. Given that within each chain it is typically the case that the direct relation is stronger than the indirect one (due to the multiplication rule), reasoners may have used this constraint as a structural heuristic for their estimates about direct and indirect causal relations (see also Experiment 2, for further tests).

Interpolation in Causal Chains

In the previous section, we have seen that in scenarios suggesting lengthening participants tend to expect a weakening of probabilistic dependencies with increased distance within the causal chain. Our main focus in the present study is another type of extension of chains that is generated by *interpolations* of mediating variables. Interpolations are frequent in contexts of causal discoveries of mediating mechanisms. Reversing the revision process in the example from the last section, we may first have obtained reliable covariation knowledge indicating that JPH3 covaries with Lipogastrosis. In a causal model, we might represent this as a direct causal relation between the two variables (see Figure 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 by another condition, that is called Hepatocitosis. On the surface, we see again a three-step chain, like the one in Figure 2B, that appears longer after the interpolation, but this effect is attributable to a very different process than the lengthening that we obtain when adding variables at the beginning

Figure 3
Example for the Extension of a Causal Chain Resulting From the Interpolation of a Variable



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 discovering 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 explicitly been addressed in the literature so far.

Weakening in Causal Interpolations

The interpolation paradigm is very different from previously investigated ones in which direct and indirect causal relations were compared. In the study by Bes et al. (2012), for example, two static three-variable chains were compared referring to different scenarios. In their paradigm it is left open how knowledge about these chains was formed. The three variables may have all been observed at once, or knowledge about the chain components may have been obtained consecutively through lengthening or interpolation.

The novel component of our new interpolation paradigm is that we highlight the temporal sequence in which knowledge has been acquired. Our paradigm consists of two consecutive temporal phases in which a belief revision process is described. In Phase 1, learners are presented with a set of trials showing the covariation between two variables, for example JPH3 and Lipogastrosis. No probability assessment is requested after this phase. In the following Phase 2, which mentions subsequent research, causal models are introduced. For example, in a direct causal condition participants are informed that, in the meantime, scientists have discovered that the two variables are linked in a direct causal relation (as in Figure 3A), whereas in the indirect or interpolated condition they are informed that, in the meantime, knowledge about a mediating mechanism has been obtained and that scientists have discovered that the disease Hepatocitosis is part of a mechanism that links JPH3 and Lipogastrosis (see Figure 3B).

Note that, unlike in previous research comparing direct and indirect causal relations, this paradigm clearly states a sequence in which knowledge has been acquired. It is explicitly stated that Hepatocitosis is a variable that has been later discovered as being part of the mechanism linking JPH3 and Lipogastrosis. Thus, the task describes a clear case of interpolation. It is also notable that no new data are shown in Phase 2. All participants only observe the initially presented covariation between two variables (JPH3, Lipogastrosis) and are then provided with verbal descriptions of the causal models scientists have discovered in further research.

Our key question is whether participants who are asked at the end of the experiment to assess the probability linking JPH3 and Lipogastrosis would offer systematically different estimates in the two contrasted conditions despite identical learning. To preview our results, we observed a stable *weakening* effect in causal interpolations. More specifically, although all participants received identical learning data, those who were informed that the causal relation between JPH3 and Lipogastrosis is in fact mediated by the variable Hepatocitosis tended to offer weaker probabilistic estimates than those who were informed that the causal relation is in fact direct.

Interestingly, we have discussed our task with numerous colleagues from psychology, philosophy, and computer science, and many had the initial reaction that such a weakening is perfectly rational. Later, on reflection, many colleagues changed their minds, though. We found these initial intuitions by experts and laypeople puzzling and wanted to explore how they can be explained. Although our main goal in the present research was to test a descriptive theory of the weakening effect, we were also interested in the question whether this effect represents a novel, previously unknown *bias* or whether it can in fact be defended as rational.

Before we present our normative and descriptive accounts of interpolations, a possible concern with the two-phase belief revision design of our paradigm needs to be addressed. An initial reaction to the weakening effect may be that possibly *demand characteristics* may have been involved. Participants may initially estimate the contingency in Phase 1 but then, confronted with Phase 2, conclude that an alteration of their estimate is requested. We implemented several safeguards against this possibility. First, we only requested probability estimates once, at the end of the experiment, not twice. Second, we kept constant that both in the direct and the interpolated conditions participants were equally informed that research had found out more about the causal relations. Thus, if the causal discoveries mentioned in Phase 2 were viewed as a cue for altering internal estimates, this should have equally affected both conditions, which it did not. Moreover, there is no reason to assume that such a possible demand characteristic should only lead to weakening. Strengthening would also be an alteration of an initial estimate. We will below present a theory that explains the asymmetric weakening effect as being caused by differences in the causal representation of the direct and interpolated conditions.

Is the Weakening Effect Rational in Causal Interpolations?

It is an interesting question whether weakening is a bias or fallacy, or whether it can be defended as a rational response. In the past decades, a number of phenomena that initially were interpreted as irrational were reinterpreted as results of rational processes (e.g., Crupi et al., 2009; Kareev, 2000; Oaksford & Chater, 1994). Can the weakening effect be similarly explained as a rational response?

The Paradox of Knowing More

One argument showing that a weakening effect cannot be rational in causal interpolation refers to a 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 for cases of causal discovery. An increase of knowledge should not lead to increased unpredictability of the world. In the following analyses we will take a closer look at properties of causal discovery.

Causal Model Analysis

A helpful strategy to address causal inferences involves separating between causal models that encode assumptions about functional relations between causes and effects and statistical evidence. This separation follows from the approach of Pearl (2000), who has in the last decade criticized his initial view (Pearl, 1988) that causal models can be reduced to probabilistic relations (see also Pearl & Mackenzie, 2018). His more recent view claims that it is necessary to separate assumptions about functional causal models from statistical information that can be used as evidence. A similar separation between causal model representations that incorporate domain knowledge and learning data has also been proposed in psychological research on causal model theory (e.g., Griffiths & Tenenbaum, 2009; Lagnado et al., 2007; Waldmann & Holyoak, 1992; see Waldmann, 2017, for overviews).

From a causal model perspective, the *discovery* of mechanisms mediating between a target cause and effect should not *prima facie* alter the strength of their causal relation. Again, using the example of aspirin and headache, a plausible assumption is that these two events have already been linked by a mechanism before it has been discovered. The discovery of the role of prostaglandins, for example, did not insert a new variable into the mechanism, prostaglandins were already part of the mechanism prior to scientists finding out about their role. Because prostaglandins mediated the causal relation between aspirin and headache all along, their discovery should not alter the previously assumed causal strength between aspirin and headache.

A graphical demonstration of why interpolations should not change probabilistic relations 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 Hepatocitosis mediates this relationship (see Figure 4B). Where was Hepatocitosis at t1? Again, a natural assumption is that Hepatocitosis already mediated the relation between JPH3 and Lipogastrosis at t1. This mediation relation was just unknown at t1. In fact, if we assume that in the unknown underlying Bayes net (let's call it God's Bayes Net), there are an infinite number of mediating variables between JPH3 and Lipogastrosis that await discovery (Figure 4B shows a fragment), then further discoveries (e.g., Figure 4C) should also not change the strength of the relation between the initially discovered two variables.

The claim that discoveries leave causal relations invariant is based on the assumption of *causal stability*. With causal stability we mean that in many domains, such as physics, chemistry, biology, medicine (i.e., natural kinds), our default assumption is that the underlying causal mechanisms tend to be invariant over time. Owing to causal stability, discoveries about mechanisms tend to be conceived of as referring to preexisting but unknown mechanisms; they do not create new mechanisms. For example, if we have observed a covariation between a virus or drug and a novel disease, it is plausible to assume that these causal events have already been linked by a hidden stable mechanism before the mechanism has been discovered.

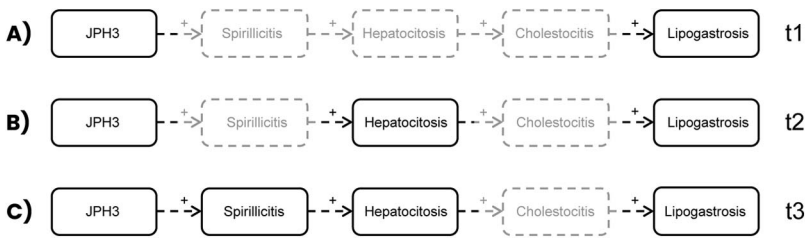
Causal stability is not a universal feature of all domains. We have different assumptions about artifacts, for example. If a company works on artifacts, such as cell phones, TVs, or cars, it is likely that the mechanisms are being altered between generations of the products. Here we should not expect stability. Therefore, no weakening effect would be predicted for artifacts. The typical goal of developments of artifacts is to strengthen the causal reliability of the device, not to weaken it (see also General Discussion). In our research, we focus on natural kind domains in which causal stability seems to be the default assumption.

Statistical Evidence

Another possible path to justify the weakening effect might focus on the fact that statistical data are typically unreliable. Probability estimates rely on samples so that a specific degree of uncertainty is always attached to these estimates. Indeed, some of the mentioned attempts to rationalize apparent biases and fallacies in other tasks focus on how uncertainty may influence participants' inferences.

In our experiments, in Phase 1, we presented participants with a limited set of observations (typically $N = 48$) of a covariation between two variables (e.g., JPH3, Lipogastrosis) and then, in Phase 2, instructed them that later scientists had found out that the two variables were either directly or indirectly causally linked. Although participants only observed one learning sample and were just once requested to provide a probability estimate, it seems plausible to assume that the scientists based their conclusions on further observations, thus increasing the sample on which the new

Figure 4
Illustration of the Process of Causal Mechanism Discovery Within God's Bayes Net



theory is based. Although the presented learning sample is not particularly small to assess a covariation between two variables, it is not very large either. Thus, a possible hypothesis might be that participants may have inferred that it is rational to assume that the probabilistic relation is weakened in the interpolation relative to the direct causal condition in the unobserved larger samples underlying the causal discoveries (Phase 2). In the following, we will discuss several possible cases that may lead to alterations of estimates with a focus on the key question whether in these cases a systematic weakening effect can be rationally defended.

Case 1: Statistical Uncertainty

Statistical uncertainty owing to the limited observed learning sample may be one factor underlying alterations of an initial internal estimate. However, there is no reason to expect that this should asymmetrically lead to lower probability estimates in the indirect interpolation conditions than the direct cause conditions (i.e., weakening), which we observed in all our studies. For one, to prevent ceiling or floor effects, we used covariations that were sufficiently far away from the extremes 1 and 0. If uncertainty was due to unreliability of the measurement, a symmetric confidence interval would be expected. Also, given that both contrasted conditions equally suggested that additional samples had been observed in the course of the discovery of the direct or indirect causal relations, a rational inference would be in both conditions to place a confidence interval around the estimated mean, which covers both weakening and strengthening.

Uncertainty could also be construed as a tendency to combine the observed data with a prior about causal strength participants may bring to bear on the task. 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 et al., 2008; Mayrhofer & Waldmann, 2016; Yeung & Griffiths, 2015).

A third possibility would be that participants may tend to provide conservative estimates when confronted with a sample size increase with unknown properties (see Rottman & Hastie, 2014, for a review of this factor). However, again, there is no reason why conservativism should affect the direct and interpolated conditions differentially. Participants in both conditions were instructed about discoveries in Phase 2.

Case 2: Context Changes

Changes of observed probabilistic relations can, even when the underlying causal model is assumed to be stable, also be expected when new evidence is collected in a changed causal context (see Cheng & Lu, 2017; Pearl & Bareinboim, 2014). For example, the relation between a drug and hypertension could be either observed in a hospital specialized on heart disease or in a group of healthy young students. Although causal strength, which is an unobserved parameter, should be unaffected by these context changes because of the stability assumption (Cheng & Lu, 2017), the observed probabilities will differ as a result of the different strengths of alternative causes. Research has shown that learners can disentangle causal strength or power from such context factors in experimental paradigms that made specific context changes highly salient in the instructions (Liljeholm & Cheng, 2007). In our studies the instructions did not mention any systematic context changes so that there was no reason for participants to expect specific context changes.

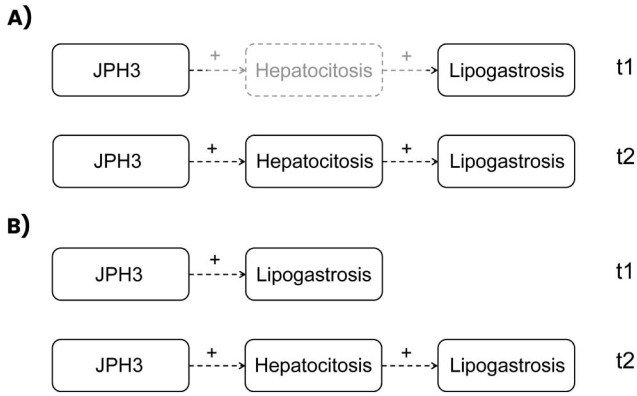
But even if context changes were expected, there is no reason to asymmetrically favor weakening over strengthening. Possible factors leading to alterations of observed probabilistic relations in new contexts include additional alternative causes, disablers, or enablers. For example, an added disabler may lower a probabilistic relation, but the addition of a previously absent enabler may strengthen it. Alternative causes may also lead to changes in both directions depending on whether they exert generative or inhibitory influences on the target effect. Moreover, context changes potentially can affect both conditions, the direct and the interpolated condition, so that no difference between the conditions should be expected.

A further possibility may be that occasionally causal mechanisms may be different in different populations. Aspirin may have different effects in subsets of the population depending on genetic differences. But even if cases can be found in which a target cause and effect are the same, there is no reason to believe that a new mixture of populations in a different sample would systematically lead to weakening but not strengthening or invariance. Again, alterations in both directions seem conceivable in both contrasted conditions. In sum, although context changes may lead to changes, they do not systematically favor weakening. Moreover, none of these possible context changes was highlighted in the instructions.

Case 3: Mechanism Changes

So far, we have discussed situations in which a rational response to uncertainty should lead to a symmetric confidence interval in both contrasted conditions, not just weakening. However, there is one possible assumption that would make it indeed plausible to systematically weaken the estimates. In the causal model section we have argued that under the assumption that causal stability holds in our learning domains, the discovered mechanism should be understood as being part of the mechanism all along even though its existence and role has been unknown prior to its discovery. This situation is shown for the example of JPH3, Hepacitosis, and Lipogastrosis again in Figure 5A. Given that

Figure 5
Two Alternative Cases of Belief Revision: Interpolation of a Preexisting Mechanism Variable (A) Versus Addition of a New Variable (B)



nothing has changed in the underlying mechanism between t1 and t2, no changes should be expected just based on the underlying causal structure. Hepatocitosis was equally operative both prior and after its discovery. However, in Figure 5B a different possibility is depicted. Here, the discovered variable Hepatocitosis has been added to the mechanism at t2 but is assumed to have been absent prior to its discovery at t1. Under the assumption that a new variable has been added to a previously simpler causal structure, it is plausible to assume that the new variable comes with enablers, disablers and alternative causes that may introduce various possibilities of disruption of the causal process. These disruptions might indeed entail a weakening of the JPH3 and Lipogastrosis relation relative to the structure at t1 in which the new variable was absent.

An important difference between the two belief revision processes depicted in Figure 5A and 5B is that Figure 5A shows a process in which a *preexisting* mechanism is *discovered*, whereas in Figure 5B the new mechanism variable is *added* to a simpler direct causal relation mechanism. Superficially, both cases involve a transition from a two-variable relation (t1) to a three-variable relation (t2). By just looking at t2 in both cases, the two cases look identical. To appreciate the difference it is again necessary to separate the causal model level from the evidence level, following Pearl's (2000) strategy. On the causal model level, the difference between discovery and insertion of a variable is whether Hepacitosis, whose role was unknown at t1 in both conditions, is retroactively inserted into the two-variable model at t1 (Figure 5A) or whether it is treated as absent at t1 (Figure 5B).

Implications for the Causal Strength Parameters

Once the two cases are differentiated on the causal model level, the question arises whether they have different implications for the causal strength parameters linking the variables. On the surface, both cases involve a transition from a two-variable model to a three-variable model. The traditional assumption underlying Bayes nets is that they are *modular* (i.e., they can be arbitrarily extended by adding new variables and new links). A plausible assumption within this approach is that people have priors about causal strength which are updated by data. Thus, one possible way to represent the transition between a two variable to a three variable model may be that people may attach similar causal strength priors to the two new links at t2 (JPH3→Hepacitosis→Lipogastrosis) as at t1. If the priors are probabilistic, a weakening effect would be entailed by the multiplication rule (Equation 1). Thus, participants may generally use the heuristic that direct relations are stronger than indirect ones. In Experiment 2 we will test this account of the weakening effect in interpolation tasks.

Our question here is whether this modular account of weakening is normative. One problem of this modular representation of the transition from a two- to a three-variable model is that it does not capture the crucial distinction between the case in which a preexisting variable has been discovered (Figure 5A) and the case in which the discovered variable is new (Figure 5B). Thus, it misrepresents interpolations. To capture the difference between these two cases a *holistic* representation of causal strength estimations is required that takes into account the task and the domain under investigation.[1]

The difference between discoveries of preexisting variables and discoveries of new variables has in fact implications for both the induction of the causal strengths of individual links and it places mutual constraints on the sizes of the different links within the interpolated model. For example, in Figure 5A the mechanism is preexisting, which implies that the observed probability linking JPH3 and Lipogastrosis is at both t1 and t2 generated by the preexisting mechanism including the yet unknown mediating variables. If, for example, the contingency linking JPH3 and Lipogastrosis is 0.42, then the links connecting the newly discovered variable Hepatocitosis should on average have a contingency of 0.65. Assuming no substantial changes between the causal contexts at t1 and t2, this contingency should hold at both t1, in which Hepatocitosis was unknown, and at t2, in which it has been discovered. Moreover, because of the multiplication rule (Equation 1), the more variables mediating the JPH3-Lipogastrosis relation have been discovered, the higher the average contingency of each link should be. This increase guarantees that the contingency between JPH3 and Lipogastrosis stays invariant despite the additional variables in the underlying causal model. Moreover, increasing the average strengths of the newly discovered links counteracts potential weakening effects of additional variables and their attached disablers, enablers, or alternative causes. Thus, in interpolations causal strength estimation normatively should be a holistic process influenced by assumptions about the domain (e.g., causal stability, discovery of preexisting variables) and the number of interpolated variables.

By contrast, the process depicted in Figure 5B does not entail such a systematic holistic constraint on causal strength. If, for example, a new variable is added to a preexisting chain, there is no reason to adapt the new causal links to the length of the chain as in interpolations. It is true that the links connecting the added variable (e.g., Hepacitosis) with the rest of the chain are in principle unconstrained, but a plausible strategy would be to either use a causal strength prior (as in the modular approach) or adapt the link strength to the strengths of the other links or the strength of the covariation learned in Phase 1. When causal strength parameters are not adapted to the length of the chain, weakening should indeed be expected when previously absent variables with probabilistic links are added.

We believe that Figure 5B may indeed represent what participants showing a weakening effect assume in the context of interpolations. But is it rational? Under the causal stability assumption, the answer is clearly no. Discoveries of preexisting mechanisms do not create them, they just make them known. We will further discuss the plausibility of this assumption for natural kind domains in the General Discussion.

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 can explain it. Thus, the main focus of our research will be on testing psychological factors that might underlie the weakening effect in causal interpolations.

Our psychological account is closely modeled after the normative analysis of the interpolation paradigm. We have outlined one condition in which it seems reasonable to expect weakening. If, as

[1] We thank Bob Rehder for suggesting the distinction between a modular and a holistic account of causal strength estimation.

depicted in Figure 5B, participants in the interpolation condition view the discovery of a mediating variable (Hepatocitosis) in Phase 2 as an addition of a previously absent variable, it would indeed be reasonable to expect a weakening relative to the direct condition in which no further variable was added.

One interesting question is why participants' inferences may violate the causal stability assumption and rather be consistent with a change of the underlying causal model (Figure 5B). One possibility may be that participants actually believe that discoveries generally extend the underlying causal model instead of identifying preexisting variables. Under the causal stability assumption, it seems very unlikely that participants generally believe in instability and thus misrepresent interpolations as insertions of variables instead of as discoveries of preexisting variables. In fact, when presenting the figure showing God's Bayes net that graphically highlights the difference between known and unknown mechanism variables (see Figure 4), most colleagues agreed that the picture correctly represents what is going on in discoveries of preexisting mechanisms. Also scientific inferences seem to be consistent with stability assumptions. Current discoveries are typically inferred backward to past cases for which no explanation has been available (see General Discussion, for a further discussion of the stability assumption).

However, there is an alternative psychological hypothesis which attributes the effect to superficial processing of the task. In the interpolation paradigm there is a shift from a structure with two variables in Phase 1 to one with three in the interpolation condition in Phase 2. Understanding this as a case of a discovery of a preexisting variable as in Figure 5A would require participants to retroactively infer a mediating variable in Phase 1 (Hepatocitosis) in which it was unknown. Such retroactive inferences are computationally demanding compared with just going with the simpler heuristic that variables that are not mentioned are probably absent.

Evidence for this hypothesis comes from studies about diagnostic blocking (Waldmann, 2000, 2001). In a diagnostic blocking task, learners observe in Phase 1, for example, that a new blood substance, Substance 1, is deterministically caused by a new disease, Midosis. Then in a subsequent Phase 2, participants are instructed that a second substance, Substance 2, that had not been measured in Phase 1, has been discovered as a potential second causal indicator of Midosis. Thus, in Phase 2 participants observe trials showing that Midosis deterministically causes both substances. Under the causal stability assumption participants should retroactively infer that the second substance had already been present in Phase 1, it just was not measured. Nevertheless, in some experiments a small tendency was observed to give lower predictiveness ratings to Substance 2 than Substance 1, which is consistent with the belief that Substance 2 may have been absent in Phase 1. However, additional studies showed that this inference is not based on a belief in causal instability but rather a manifestation of attempts to simplify processing effort. In simplified salient tasks this effect disappeared (see López et al., 2005; Waldmann & Walker, 2005). Thus, our hypothesis in the interpolation task is that participants' inferences are not triggered by beliefs in causal instability, but rather are a consequence of attempts to reduce processing effort.

One processing mechanism that might provide a psychological explanation would be the assumption that participants may use mental simulations as a technique to arrive at inductive inferences when additional information is provided (see Griffiths et al., 2012).[2] In the interpolation conditions they may first learn the two-variable covariation. After being instructed about the discovered mediating variable, without being shown new data, they may simulate the augmented causal model in Phase 2 while failing to adapt the causal strength parameters to the interpolation situation. This process would lead to weakening. A proper understanding of interpolations would on this account require to retroactively simulate the trials of Phase 1 with the proper causal strength parameters after the instructions for Phase 2 have been provided, which is obviously hard.

To sum up, we have outlined three possible accounts of the interpolation task. On a normative account, no weakening effect is expected when a preexisting mechanism is later discovered. Moreover, causal strength parameters should be induced for the newly discovered links that preserve the observed covariation between the variables they are mediating. The second account, which is consistent with a modular or heuristic account, predicts that learners generally assume as a heuristic that indirect relations are weaker than direct ones. This account is insensitive to the order in which direct and indirect relations are acquired and it does not distinguish between a case in which a preexisting variable is discovered versus a new variable is added to a previous simpler causal model. Finally, the third account of weakening in interpolations, our belief revision account, is that learners tend to have difficulties with retroactively considering that the newly discovered variables had already been part of the mechanism prior to their discovery. These difficulties lead to effectively treating newly discovered mechanism variables as previously absent although domain knowledge suggests their presence in the past.

Preview of Experiments

The presentation of the experimental series starts with two experiments in which we set the stage for later studies by demonstrating the existence of the weakening effect presenting variables on different levels of description (Experiments 1a, b). The next two experiments, which together tested 1,980 participants, investigate boundary conditions of the weakening effect (Experiments 2a and 2b). Whereas two conditions correspond to the standard interpolation paradigm with a covariation learning phase presenting two variables (Phase 1) preceding a causal model instruction phase (Phase 2) in which either a direct or an indirect interpolation condition were instructed, in the two further conditions the sequence was reversed. Here the causal models were instructed first, followed by covariation learning. This reversal was motivated by our experience that the presentation of God's Bayes net (see Figure 4) convinced most of our audience that a weakening effect is biased because the figure provides a salient presentation of the fact that the mechanism was already in place prior to discovering its components. We expected that the reversal should make the weakening effect disappear because now learners can in both conditions map the respective causal model on the learning trials. Both models, the direct and the indirect causal model, can be parameterized in a way to be consistent with the learned contingencies.

The two experiments contain further tests of our theory. As elaborated above, an alternative hypothesis that ignores the belief

[2] We thank Mike Oaksford for this suggestion.

revision component altogether might claim that participants may merely use the heuristic that indirect causal relations are generally weaker than direct ones. This hypothesis falls out of a modular account of Bayes nets, which assumes that learners generally use causal strength priors for individual links regardless of whether the task involves causal interpolations of preexisting variables or discoveries of new variables. On this account, these domain distinctions do not matter, the key feature triggering weakening is the contrast between direct and indirect relations. Thus, the sequence of presenting direct relations and chains should not matter on this account. Bes et al.'s (2012) study provides evidence that seems to support this hypothesis. However, as argued in the introduction, their effect may have been due to the fact that in their studies direct relations (e.g., A-B in A→B→C) were always presented as components of indirect ones (e.g., A-C in A→B→C) so that this heuristic seems reasonable there. By contrast, our Experiments 2a and 2b employ a between-subjects design in which the direct relation is learned separately from the indirect one. If participants used the heuristic that indirect relations are weaker than direct ones here as well, a weakening effect should be seen in both the interpolation condition and the new conditions in which causal models were instructed prior to learning. Thus, Experiment 2 additionally provides a test between a simple heuristic based on a modular understanding of causal Bayes nets and our belief revision account.

A third change involves the test question. In all our experiments we asked participants to assess the conditional probability of the effect (e.g., Lipogastrosis) given the cause (e.g., JPH3) without further describing the sample the test subjects were drawn from. By contrast, in Experiments 2a and 2b we explicitly informed participants that the randomly selected test subject for which an estimate should be made comes from the learning sample. This way we rule out that participants consider possible context and sample changes between Phase 1 and Phase 2. If these were the crucial factors underlying the weakening effect, the effect should disappear.

Experiments 3 and 4 investigate two interrelated predictions of our theory. Experiment 3 tests the hypothesis entailed by our theory that causal interpolations lead to longer chains with more variables, which potentially highlights multiple possibilities of how things can "go wrong." Participants might be led to consider enablers, disablers or alternative causes affecting the mediating variables when the chain contains more variables. 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. Under the assumption that participants represent the new variables in the indirect condition in Phase 2 as having been absent in Phase 1 (see Figure 5B), the added variables along with their disablers, enablers or alternative causes should lead participants to consider disruptions of the causal process which would predict a weakening effect. We focus in Experiment 3 on disablers as an example for this hypothesis and test it by manipulating whether disablers are explicitly mentioned or not (i.e., explicit vs. implicit representations).

Experiment 4 complements Experiment 3. Adding variables to a given causal relation should only lead to a weakening effect if these variables are newly inserted and have not already mediated the observed covariation prior to the discovery of the mechanism. If learners had a correct understanding of interpolation, they should understand that the discovered variables along with their disablers, enablers, and alternative causes already mediated the covariation in Phase 1 prior to scientists having found out about them (see Figure 5A). On our normative holistic account of causal strength estimation discoveries of preexisting variables have implications for the causal strength parameters of individual links as a function of the number of discovered variables. The more mediating variables are discovered, the stronger the causal strength parameters should become in a causal chain. This relation is entailed by the multiplication rule (Equation 1). A systematic strengthening of the assumed link strengths, thus weakening the influence of disablers, ensures the invariance of the covariation learned in Phase 1. Experiment 4 tests whether learners have an understanding of this relation between causal strength and the length of a chain in interpolation tasks. More specifically, if they had an adequate understanding of the implications of interpolations, we should see a systematic increase of strength estimates with increased length (Figure 4, Figure 5A). Otherwise, if, as we hypothesize, participants misrepresent interpolations as situations in which new variables are added to a simpler causal model, no such trend should be observed. In sum, both Experiments 3 and 4 target different, but strongly interrelated implications of our weakening account.

Most of our experiments study chains as an example for an indirect causal relation. How about interpolating more complex network structures between two variables (cf. Figure 1)? If the network discovered in Phase 2 is assumed to be preexisting but unknown in Phase 1, no weakening should normatively be observed. Regardless of the complexity and the parameters connecting the mediating variables, the causal contingency between the two learning variables should not systematically change for the same reasons why we do not normatively expect such a change in simple chains. However, we expect that participants will again misrepresent the discovery in Phase 2 as a situation in which new interconnected variables are added that had not been present in Phase 1. If that was the case, the structure and the parameters of the inserted causal network should again influence the inferences about the two target variables. For all our experiments, the experimental materials (including example video clips) and data can be accessed under https://osf.io/aqzps/ (Stephan et al., 2020).

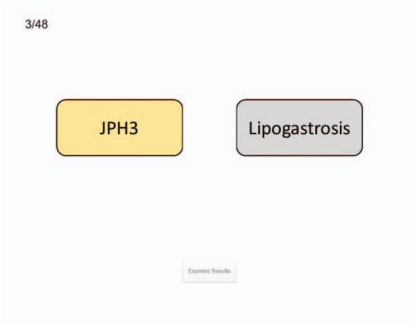
Experiment 1a

The goal of Experiment 1 was to set the stage for our project by testing whether we will find a weakening effect after causal interpolations. The focus of our experiments is on natural kind domains (such as physics, chemistry, biology, medicine) for which it is plausible to assume causal stability. We will discuss other possible domains in the General Discussion. To control for effects of prior knowledge, we decided to employ fairly abstract materials. Therefore, the causal variables referred to unknown variables of a fictitious biological scenario.

Experiment 1a employed the basic two-phase belief revision paradigm that we used in all studies. In the first phase (the contingency-learning phase, Phase 1), all participants were presented with trial-by-trial learning information about the values of two variables. This learning phase allowed participants to acquire

knowledge about the degree of covariation between the two variables. Subsequently in Phase 2, participants were informed that scientists had later discovered that these two variables were either directly or indirectly causally related. The causal model information was manipulated between subjects. In the direct causal condition, participants were told that the two variables observed in Phase 1 were in fact directly causally related. In the indirect interpolation condition, participants were instructed that scientists had found that a new mechanism variable mediates the previously observed covariation. Thus, the indirect interpolation condition implements a case of causal discovery of a previously unknown mechanism in which a new variable is interpolated between the two variables observed in Phase 1. No new data were shown in this second phase. Then, the final test question requested participants to estimate the conditional probability linking the two variables. A weakening effect is observed when the probability estimates are lower in the indirect (interpolated) causal than in the direct causal condition, despite identical learning data.

Figure 6
Example of a Case From the Trial-by-Trial Learning Task Used in Experiment 1



Note. The illustration depicts a case in which the cause was present (yellow JPH3 box) but the effect was absent (gray Lipogastrosis box). See the online article for the color version of this figure.

Method

Participants

One hundred forty participants (100 female, 39 male, one participant indicated to be neither male nor female, M_{Age} = 36.69, SD_{Age} = 12.50) recruited via Prolific (www.prolific.ac) participated in this online study and provided valid data. The rationale behind this sample size was that it allowed us to detect a medium effect size of d = 0.50 with more than 80% probability. The inclusion criteria were a minimum age of 18 years, English as native language, at least an A-level degree, and an approval rate concerning participation in previous studies of 90%. Participants were asked to participate only via laptop or desktop computer and not via smartphone or tablet, because we wanted to minimize the chances of distractions (e.g., public places, subway). Participants received a monetary compensation of £0.70 for their participation.

Design, Materials, Procedure

The study employed a between-subjects design (causal model: direct causal condition vs. indirect [interpolated] causal condition). An example video illustrating the experimental procedure in one of the conditions can be viewed at https://osf.io/evusj/. As a cover story we employed a fictitious scenario according to which biologists were interested in studying the statistical relation between the mutation of the gene *JPH3* (J) and the gastro-intestinal disease *Lipogastrosis* (L), which was defined as characterized by an excessive accumulation of fats in the digestive tract. Participants were further instructed that the biologists had conducted a study in which they examined two random samples of mice for the disease, one in which individuals were carrying the mutation and a second one in which they were not. Participants were informed that the results of the biologists' study would be presented to them serially via a graphical display (for more details see below and Figure 6) and that their task was to examine the results thoroughly, without taking any notes. To proceed to the learning task, participants had to pass an instruction check involving two multiple-choice questions referring to the hypothesis of the biologists and the meaning of the colors of the nodes which coded whether the JPH3 mutation and/or Lipogastrosis were present or absent. The contingency that we presented to participants in the subsequent learning task is shown in Table 1. The probability of Lipogastrosis given a JPH3 mutation, $P(L|J)$, was .75 and the probability of Lipogastrosis in the absence of JPH3 mutation, $P(L|\neg J)$, was .21. Hence, the contingency $\Delta P_{JPH3\text{-}L}$ was .54.

In the learning phase, we used a trial-by-trial observational learning task in which the 48 cases summarized in Table 1 were presented to participants in random order on a computer screen. An example of what the screen looked like during the learning phase is shown in Figure 6. An example video of the learning task can be accessed via https://osf.io/8cjft/. The presence of either variable (J or L) was indicated by a yellow text box, while their absence (¬J or ¬L) was indicated by a gray text box. 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.

After the learning task, participants were given information about the causal model assumed to underlie the observed relation between JPH3 and Lipogastrosis. Depending on condition, participants were either instructed that JPH3 and Lipogastrosis were *directly* or *indirectly* (by means of an interpolated variable) causally related. In the direct causal condition, participants were presented with the following text along with the illustration shown in Figure 7A.

> 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.

Participants in the indirect (interpolated) causal condition were presented with the following text together with the illustration shown in Figure 7B.

> 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. Specifi-

Table 1
Contingency Presented to Participants in Experiment 1a

Observations				Conditional probabilities		Contingency
n(J, L)	n (J, ¬L)	n (¬J, L)	n (¬J, ¬L)	P(L|J)	P(L|¬J)	ΔP
18	6	5	19	0.75	0.21	0.54

cally, the JPH3 mutation can sometimes lead to Hepatocitosis, an abnormal occurrence of hepatic enzymes. This is indicated by the arrow (with a + sign) that goes from JPH3 to Hepatocitosis. Finally, Hepatocitosis can sometimes lead to Lipogastrosis. This is indicated by the arrow (with a + sign) that goes from Hepatocitosis to Lipogastrosis. Other factors can also influence the disease.

To control for possible associations with specific diseases (e.g., liver diseases), we used different labels for the interpolated variables. More specifically, participants either learned that the interpolated variable was *Hepatocitosis* (see Figure 7B), described as an abnormal increase in hepatic enzymes, or *Cholestocitis*, described as an abnormal occurrence of cholesterol, or *Spirillicitis*, described as an infection colonizing the gut, or *Paracelocitis*, described as a dysfunction of the process involved in fat metabolism. After having read the causal model information, participants proceeded to the test screen. They were informed 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. Participants in the direct causal condition were shown the illustration depicted in Figure 7C. Participants in the indirect (interpolated) causal condition were shown the illustration depicted in Figure 7D, with the interpolated variable being the one presented in Phase 2. Participants were asked to estimate the predictive probability of Lipogastrosis given JPH3, P(L|J). The phrasing of the test question was: "What do you think is the probability that the mouse also has Lipogastrosis?" Participants 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."

Results and Discussion

The results are summarized in Figure 8. Participants in the direct causal condition gave ratings (M = 68.47, 95% CI [64.49, 72.45]) that were close to the normative value of P(L|J) = 0.75, whereas the ratings in the indirect (interpolated) causal condition were lower (M = 56.30, 95% CI [51.15, 61.45]). An independent t test confirmed that the observed difference was significant, $t(138) = 3.73$, $p < .001$, $d = 0.63$.[3]

The results of Experiment 1a indicate that interpolating causal variables between two covarying causal variables changes reasoners' representation of the observed probabilistic relation. More specifically, although the interpolated variable was introduced as a mediator between the two variables, a "weakening effect" was found in the indirect (interpolated) condition compared with the direct causal condition. This finding is consistent with the theory that participants misrepresent interpolations and treat the newly introduced variable in Phase 2 as a new variable that is added to the causal model.

Experiment 1b

In Experiment 1a, the interpolated variable (e.g., Hepatocitosis) and the effect variable (Lipogastrosis) belonged to the same category, physiological conditions (two diseases involving dysfunctional physiological processes). Although the diseases were novel, it cannot be ruled out that participants may have the abstract intuition that two separate diseases may not be strongly probabilistically connected. We wanted therefore to make sure that the observed effect is not restricted to cases in which a separate disease is interpolated as a mediator and to test whether the weakening effect interacts with the type of variables that are being employed.

We constructed three different versions of the cover story used in Experiment 1a by varying the level of description of the three variables following the root cause. We contrasted the physiological level with a genetic and a molecular one (see Table 2). A further change was that we increased the semantic coherence of the description of the chain mechanism across the links. For example, in the genetic condition several genes were connected that led to abnormal mutations between the causally connected genes. Our goal here was to test whether the weakening effect will also be observed when both the type of linked events and the type of mechanism are very similar. Furthermore, we introduced two variables in the indirect (interpolated) causal condition instead of one to test the generality of the weakening effect.

Method

Participants

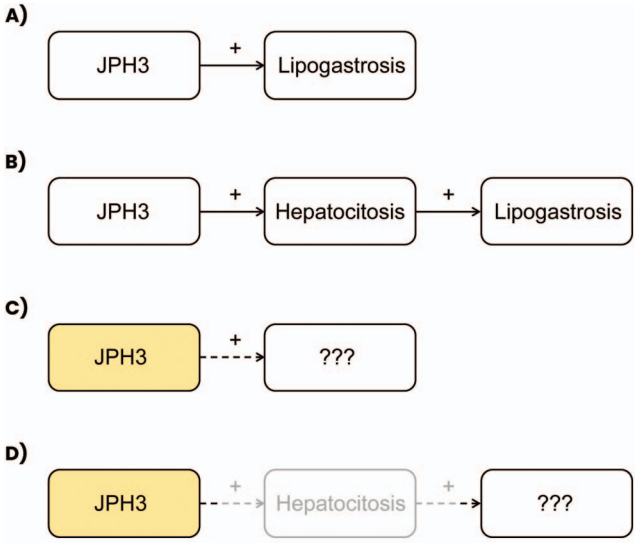
Five hundred ten new participants (242 female, 265 male, three participants indicated neither male nor female, M_{Age} = 33.29, SD_{Age} = 14.73) were recruited via *Prolific* and provided valid data. This sample size allows us to detect small main effects and interactions of $f = 0.15$ (d = 0.3) with more than 80% probability. We planned for a smaller effect this time because we were uncertain whether our new level of description variable would diminish the weakening effect. The inclusion and exclusion criteria were identical to the ones of Experiment 1a. Participants were paid £0.70 for their participation.

Design, Materials, and Procedure

The study employed a 2 (causal model: direct vs. indirect [interpolated] causal relationship) × 3 (level of description: physiological vs. genetic vs. molecular) between-subjects design. We used the same contingency data set as in Experiment 1a. An example video illustrating the experimental procedure in one of the conditions can be accessed at https://osf.io/ufche/. The cover story and the procedure were similar to those of Experiment 1a with the exception that, depending on condition, the interpolated variables and the effect variable were described either in terms of a genetic, physiological or molecular-level process. The root cause was always the genetic mutation of JPH3 (as in Experiment 1a). Unlike in Experiment 1a, we instructed indirect chains with two interpolated variables.

[3] A meta-analytic summary of the sizes of the weakening effects is provided in the General Discussion.

Figure 7
Illustration of the Causal Models Shown to Participants in Experiment 1a



Note. A and B show the causal models that participants were shown after the learning data in the direct and the indirect (interpolated) causal conditions, respectively. C and D show the causal model illustrations participants were shown on the test query screen. In the indirect (interpolated) causal conditions, the interpolated variable was either Hepatocitosis, Cholestocitis, Spirillicitis, or Paracelocitis. See the online article for the color version of this figure.

As in Experiment 1a, participants first read that a group of biologists was interested in 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 "an abnormal mutation of a gene in the digestive tract" (genetic-level condition) or as "an abnormal modification of fats in the digestive tract" (physiological-level condition) or as "an abnormal change of the molecular structure of enzymes in the digestive tract" (molecular-level condition). As in Experiment 1a, participants were first presented with the contingency data in a serial trial-by-trial learning task (Phase 1), and then were given the causal model information using an illustration and a short text (Phase 2). For example, in the direct causal condition, participants were given the following description:

> 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 (which depending on the condition might be an abnormal gene modification, an abnormal modification of fat, or a change of the molecular structure of enzymes).

In the indirect (interpolated) causal conditions participants were given the following causal model information:

> 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 [the information in the square brackets varied according to condition Table 2].

After participants had read the causal model instructions, they proceeded to the test scenario, which was analogous to the one in Experiment 1a. Participants were generally asked to estimate the probability of LipoX given JPH3.

Results and Discussion

The results are summarized in Table 3. We found that participants tended to give lower ratings in the indirect (interpolated) causal conditions than in the direct causal condition, irrespective of the level of description and type of variable. A 2 (causal model: direct vs. indirect [interpolated] causal relationship) × 3 (level of description: physiological vs. genetic vs. molecular) factorial ANOVA revealed a significant main effect of "causal model," $F(1, 504) = 20.78$, $p < .001$, $f = 0.203$ ($d = .41$), confirming again a weakening effect. There was, by contrast, no effect of "level of description," $F(2, 504) = 1.10$, $p = .33$, and also no interaction between "causal model" and "level of description," $F(2, 504) = 0.37$, $p = .69$.

The results of this experiment show that the weakening effect is robust and does not depend on the level of description and the degree of semantic coherence of the instructed causal mechanism.

Experiment 2a

After having established the weakening effect with different types of variables, Experiment 2a[4] tests some of the assumptions underlying our belief revision theory of the weakening effect. The experiments also serve as a test against the alternative modular heuristic account of the weakening effect.

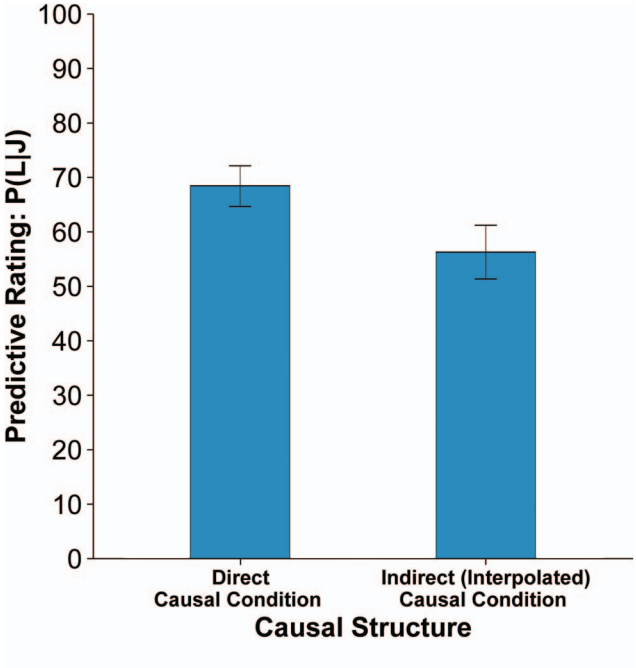
As control conditions we ran again the two-phase interpolation paradigm in which after a covariation learning phase (Phase 1) either a direct or an indirect causal relation with an interpolated variable was instructed. As in Experiment 1, no further data were presented and the test question requesting a conditional probability estimate was asked only once after Phase 2.

Our theory predicts a weakening effect because learners in the interpolation condition may have difficulties with retroactively considering that the newly discovered variable had already been part of the mechanism prior to its discovery. According to the simulation account proposed in the introduction participants should have difficulties with retroactively simulating the learning trials with the full causal model. It should be easier to represent Phase 1 in terms of a two-variable model in both conditions and add a mechanism variable in the indirect interpolation condition (Phase 2).

As a test of this hypothesis we added two conditions in which the sequence of phases was reversed (causal model-first conditions). Here, participants were first instructed about either a direct two-

[4] Experiments 2a and 2b were the last experiments that we ran in the present set of studies.

Figure 8
Results (Means and 95% CIs) of Experiment 1a



Note. See the online article for the color version of this figure.

variable (JPH3→Lipogastrosis) or an indirect three-variable causal chain (JPH3→Hepatocitis →Lipogastrostis) in Phase 1, followed by a covariation learning phase in Phase 2 in which only the two variables that were part of both models were presented. Our prediction for the causal model-first conditions was that knowledge about the underlying causal model acquired in Phase 1 should allow learners in Phase 2 to acquire causal model parameters that are consistent with the instructed causal model and reflect the presented covariation. There is no need to retroactively simulate the trials with an added variable, the mechanism variable is already known prior to Phase 2 covariation learning. Moreover, the two phases in the causal model-first conditions do not suggest any structural changes across phases. The instructions simply stated that in Phase 2 the biologists were only interested in exploring the covariation between two variables. There is no reason to assume that a previously instructed mediating variable (Hepatocitosis) simply disappears because it was not measured in Phase 2.

The causal model-first conditions also allow us to test the alternative modular heuristic account that learners may generally ignore the two phases in belief revision and simply use the heuristic that the two outer variables in a chain generally tend to be correlated more weakly than two directly related variables. The results of Bes et al. (2012) are consistent with this hypothesis but their procedure may have prompted participants to use this heuristic. Bes et al. (2012) generally presented direct relations as components of indirect ones. By contrast, we compared in the causal model-first conditions a direct with an indirect causal relation in a between-subjects design followed by identical learning trials. Thus, our experiment does not present the direct relation as a subcomponent of an indirect one. Nevertheless, if participants used the heuristic that indirect relations tend to be generally weaker than direct ones, a weakening effect should be found in both the interpolation and the causal model-first conditions. By contrast, if we observed weakening only in the interpolation conditions but not in the causal model-first conditions, this would provide evidence for the hypothesis that weakening is a consequence of the temporally ordered belief revision process.

A further modification we implemented in the experiment involves the test question. In Experiment 1a, we asked participants to assess the conditional probability of the effect (e.g., Lipogastrosis) given the cause (e.g., JPH3) in a randomly sampled test mouse. We did not explicitly say how the test mouse was selected. Because one concern may be that the interpolation paradigm suggests that the sample was increased subsequent to the learning phase, we wanted to be unambiguous in the present study and therefore instructed participants that the test mouse comes from the original learning sample. Thus, possible context or sample changes should not affect the assessment here.

Method

Participants

Six hundred twenty participants (341 female, 278, one participant indicated to be neither male nor female, M_{Age} = 36.43, SD_{Age} = 12.59) recruited via *Prolific* participated in this online study and provided valid data. The inclusion and exclusion criteria were similar to Experiments 1a and 1b. Participants received a monetary compensation of £0.75 for their participation. Our theory

Table 2
Descriptions of the Interpolated Variables in the Indirect (Interpolated) Causal Conditions of Experiment 1b

Condition	Description
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."
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."
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."

Table 3
Summary of the Results in Experiment 1b

	Direct causal condition			Indirect (interpolated) causal condition		
Descriptive statistics	Genetic level	Physiological level	Molecular level	Genetic level	Physiological level	Molecular level
M	68.25	68.66	66.28	62.34	59.18	58.06
SD	17.92	16.63	16.47	21.99	21.15	21.88
Mdn	71.00	71.00	68.00	67.00	61.00	63.00
95% CI	[64.38, 72.12]	[65.07, 72.25]	[62.73, 69.83]	[57.60, 67.08]	[54.62, 63.74]	[53.34, 62.78]

predicts a very specific interaction pattern with a weakening effect in the interpolation conditions and the absence a weakening effect in the causal model-first conditions, which is the reason why we had to test a fairly large sample. The rationale behind the planned sample size was that it allows us to detect a small interaction effect of $f = 0.135$ ($d = 0.27$) with more than 90% probability. The sample size calculation was based on a simulation in which we assumed a medium weakening effect of about $d = 0.5$ in the interpolation conditions (with the assumed means and standard deviations being $M = 68$, $SD = 17$ vs. $M = 59$, $SD = 17$) and the absence of a difference ($d = 0$) in the causal model-first conditions (with the assumed means and standard deviations being $M = 68$ and $SD = 17$).

Design, Materials, and Procedure

The study design was a 2 (time of presentation of causal-model information: interpolation vs. causal model-first) × 2 (instructed causal model: direct vs. indirect causal relationship) between-subjects design. An example video illustrating the experimental procedure in one of the conditions can be viewed at https://osf.io/kme3d/. We used the same cover story about the JPH3 mutation and Lipogastrosis as in Experiment 1a. The instructions in the interpolation conditions were similar to those used in Experiment 1a. Participants in the condition in which the causal structure information was presented prior to the learning data either learned that a group of biologists had discovered that the JPH3 mutation and Lipogastrosis were directly causally related or that the JPH3 mutation and Lipogastrosis were indirectly causally related via Hepatocitosis. Participants were shown a graphical illustration of the respective causal structure and a short description that was very similar to how we presented the causal models in the interpolation conditions. Participants in all conditions were informed that they will see the results of a study conducted by biologists in which they investigated the relation between JPH3 and Lipogastrosis. We employed the same serial learning task as in Experiment 1a. Participants in the condition in which the causal model information was given prior to the learning data were informed furthermore that the status of Hepatocitosis will not be shown during the learning trials and that the Hepatocitosis text box will therefore be covered by a black mask. An illustration is shown in Figure 9.

Figure 9
Example of a Learning Trial From the Causal Model-First Conditions in the Indirect [interpolated] Condition

Note. See the online article for the color version of this figure.

We masked the intermediate node during the learning phase because we wanted to remind participants of the mechanism variable linking JPH3 and Lipogastrosis while at the same time ensuring that learning only focused on two variables as in the interpolation conditions. At the end, participants in both the interpolation and the causal model-first conditions proceeded to the test question, which again requested an assessment of the predictive probability of Lipogastrosis given JPH3, *P*(L|J). Unlike in Experiment 1a, however, the test question in this experiment referred to a mouse that was drawn randomly from the *original* sample learners had seen during the covariation learning phase. Also, unlike in Experiment 1, the image that participants were shown on the test query screen in both the direct and indirect conditions only showed the JPH3 and the Lipogastrosis nodes, not the mediating variable.

Results and Discussion

The results are summarized in Table 4. We replicated the weakening effect in the interpolation conditions again: Participants who were instructed that the two variables shown in Phase 1 were indirectly causally related provided lower estimates than participants in the direct condition. Table 4 also shows that no weakening was observed in the causal model-first conditions.

A 2 (time of presentation of causal-model information: interpolation vs. causal model-first) × 2 (instructed causal model: direct vs. indirect causal relation) factorial ANOVA yielded a significant interaction effect, $F(1, 616) = 10.12$, $p = .016$, $f = 0.129$. Planned contrasts confirmed that this interaction was obtained because (a) we replicated the weakening effect in the interpolation conditions, $t(616) = 3.43$, $p < .001$, $d = 0.40$, while (b) the predictive probability ratings for the "direct causal relation" and the "indirect interpolated causal relation" did not reliably differ between the two groups that had learned about the causal models prior to the presentation of the contingency data, $t(616) = -1.08$, $p = .23$, $d = -0.12$.

In sum, the results provide support for our theory that presenting causal model information after learning generates a weakening effect because participants may have difficulties with retroactively inferring the presence of the mechanism variable prior to its discovery. The experiment also rules out the alternative heuristic account which predicts a general weakening effect regardless of when information about direct or indirect causal relations was acquired. In the conditions in which causal models were introduced prior to covariation learning, no weakening effect was observed. This finding shows that learners do not generally view indirect relations as weaker than direct ones.

Table 4
Results of Experiments 2a and 2b

Measure	Experiment 2a				Experiment 2b			
	Mdn	*M*	*SD*	95% CI	*Mdn*	*M*	*SD*	95% CI
Interpolation								
Direct causal relation	70.00	65.70	18.40	[62.78, 68.62]	70.00	65.78	19.97	[63.65, 67.91]
Indirect causal relation	57.00	58.50	17.36	[55.75, 61.25]	65.00	61.94	18.16	[60.00, 63.88]
Causal model-first								
Direct causal relation	67.00	62.70	18.98	[59.69, 65.71]	68.50	64.37	19.52	[62.29, 66.45]
Indirect causal relation	69.00	64.97	19.09	[61.94, 68.00]	69	64.16	19.14	[62.12, 66.20]

Finally, the modified test question did not seem to alter the findings in the interpolation conditions, which weakens the hypothesis that assumptions about sample changes underlie the weakening effect.

Experiment 2b

Experiment 2a yielded a significant interaction between "instructed causal model" and "time of presentation of causal model information" that was predicted by our theory and refutes the alternative heuristic account. The experiment replicated the weakening effect in interpolations. Moreover, it showed a novel finding that had not been reported in the literature before: the absence of a weakening effect in a between-subjects design when the direct and indirect causal models were instructed prior to covariation learning. Given the importance of this pattern of findings for our theory, we wanted to make sure that the absence of a weakening effect in the causal model-first conditions is stable. The main goal of Experiment 2b was therefore to estimate the interaction effect with even higher precision by replicating the design of Experiment 2a.

A further goal was to broaden the generality of the weakening effect by testing a different cover story with other variables. The new cover story was inspired by the aspirin example we used in the introduction. Participants in this experiment learned about the relation between a new drug and an unfamiliar disease.

Method

Participants

One thousand, three hundred sixty participants (803 female, 552 male, five participants indicated to be neither male nor female, $M_{Age} = 34.59$, $SD_{Age} = 12.08$) recruited via *Prolific* participated in this online study and provided valid data. The inclusion criteria were the same as in Experiment 2a. The rationale behind this sample size was that we wanted to increase our measurement precision substantially. The chosen sample size allowed us to detect a small interaction effect of $f = 0.09$ with at least 90% probability. In contrast to Experiment 2a, the calculation assumed a weakening effect of $d = 0.3$ in the interpolation condition, which corresponds to the lower boundary of the 95% CI of the average weakening effect we measured across the previous studies of this project (see a meta-analytic overview of effect sizes in the General Discussion).

Design, Materials, and Procedure

The study design and experimental procedure were identical to those in Experiment 2a. The only difference was that we used a cover story in which scientists investigated the relation between a newly developed drug called "Diclofan" and a disease called "Midosis" in a human sample. An example video illustrating the experimental procedure in one of the conditions can be viewed at https://osf.io/4vbgx/.

Results and Discussion

The results are summarized in the right part of Table 4. We replicated the weakening effect in the interpolation condition. Moreover, we found again that participants' predictive probability ratings were not significantly different in the causal model-first conditions. The observed pattern of ratings is thus in line with the predicted interaction effect. A planned contrast testing the predicted ordinal interaction pattern was significant, $t(1356) = 1.74$, $p = .04$ (one-tailed), $f = 0.05$, although the measured effect size was smaller than predicted. Given that we predicted the pattern of the replication, a one-tailed interaction test seems warranted. Table 4 shows that we obtained only a small, yet significant, weakening effect in the interpolation condition this time, $t(1356) = 2.61$, $p < .01$, $d = 0.20$. The predictive probability judgments did not differ in the causal model-first condition, $t(1356) = 0.15$, $p = .89$, $d = 0.01$. Thus, in both Experiments 2a and 2b we have clear evidence for the absence of a weakening effect in the causal model-first conditions and the presence of a weakening effect in the interpolation conditions (see meta-analysis in General Discussion for the CIs of the effect sizes). Both Experiments 2a and 2b therefore provide strong evidence against the heuristic account of weakening, which predicts weakening regardless of the order of presenting causal model information in our task.

Experiment 3

Overall, Experiments 1 and 2 showed that, given a specific contingency between two variables, reasoners' probability judgments $P(E|C)$ were lower when they believed that the two variables are indirectly connected via an interpolated variable than when they believed that they are directly causally connected. We showed that this weakening effect in causal interpolations occurs robustly when the causal model is instructed after observing the covariation between the two variables but disappears when the causal model is instructed prior to learning. This pattern confirms the hypothesis that in the interpolation paradigm participants treat the newly discovered mechanism variable as if it was added to the causal model after the covariation data had been collected.

Experiment 3 tests one implication of this assumption. One determinant of the weakening effect might be reasoners' tendency

to perceive indirect causal relationships as more prone to "failure" than direct ones. This assumption entails a weakening effect if the additional variable was not treated as a discovery of a preexisting mechanism but as a new variable that introduces additional sources of disturbance. Whereas a direct causal relationship may appear very stable, an indirect one with an added variable may lead people to think about possible ways the mediating variables can be affected by outside factors, 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.

One function of considering the effects of external factors in a specific case with a probabilistic causal relation is to explain why the observed contingency is not deterministic. To account for a probability of an effect lower than 1 in the presence of the cause, one can attribute the lowering to either an intrinsically low causal power, to disablers, or the lack of enablers. Alternative generative causes can be invoked to explain why effects can occur in the absence of the cause. We predict that explicitly highlighting these external influences should lead to an exaggeration of their impact. More specifically, the consideration of additional disablers in the model with an added variable should lead to a weakening effect. Moreover, exaggerating the importance of enablers for a cause producing its effect might sensitize participants to the possibility of their absence (especially in a long chain in which many enablers need to be present to guarantee the effect). As for independent alternative generative causes, they might lead to more instances of the target effect being attributed to them rather than to the observable target cause. Thus, if the goal of the reasoner is to explain why a given causal relation is probabilistic, considering external variables as an explanation and exaggerating their strengths should typically lead to a weakening effect in a situation in which variables are represented as additions as opposed to interpolations. In interpolations the possibly disrupting effects of external variables should be counteracted by adapting causal strength estimates, thus counteracting weakening. The more preexisting mechanism variable are discovered, the stronger their causal strength should on average be on a normative account. Increasing causal strength estimates lowers the potential impact of external variables (see Experiment 4).

The prediction that explicitly mentioned external causal factors might have a stronger impact than conditions in which they are not mentioned is a psychological prediction; it is not normative. From a normative point of view, it should make no difference whether external factors are made explicit or left implicit in a causal Bayes net. In the initially represented direct causal relationship, external factors are reflected in the probabilistic relations of the observed nondeterministic contingency. In direct causal relations, the causal impact of alternative causes is implicitly represented in the probability of the effect in the absence of the cause, whereas causal power, the impact of disablers, or the probability of the absence of enablers manifest themselves in a lowered probability of the effect given its cause (i.e., something else might have contributed to the absence of the effect when the cause was present).[5] When a direct causal relation is subdivided into a chain, potential external causal influences are expected to spread along it, according to Equation 1.

However, psychologically it could make a difference whether the variables that modulate the causal relationship are explicitly mentioned or not because the explicit presence of external causal variables in the chain representation can make participants more inclined to consider their potential causal influence. Indeed, research on support theory has shown that people weigh explicit hypotheses more than implicit ones (Tversky & Koehler, 1994). Another line of research demonstrating the stronger impact of explicitly compared with implicitly presented disablers on inferences comes from research on reasoning with conditionals ("If p, then q"). This research has demonstrated that inferences from p to q are influenced by both explicitly mentioned disablers (Byrne, 1989) and implicit ones (Cummins, 1995; Cummins et al., 1991). Both modes of presenting disablers affect inferences with explicitly mentioned disablers leading to stronger effects. Experiment 3 extends this work by exploring the effect of explicit and implicit disablers in the interpolation paradigm with direct and indirect (interpolated) causal relations.

Although we expect that all kinds of external variables should affect inferences, in Experiment 3 we focused on disablers, which in line with previous research should show strong effects. Disablers should be viewed as clear examples of potential disrupters. We compared causal models in which disablers were explicitly mentioned with causal models in which they remained implicit. We predicted the impact of disablers to be higher when they were explicitly mentioned than when they were only left implicit.

A second goal was to test whether the weakening effect interacts with the explicit/implicit manipulation. We predicted an independent effect under the assumption that the presence of variables along the chain in the indirect conditions may generally invite participants to consider potential influences on each of these variables. Given that an indirect (interpolated) chain consists of more variables than a direct causal relation, we would expect that participants consider more external influences in the indirect, interpolated chain compared with the direct causal relation.

A further goal of Experiment 3 was to test whether reasoners solely consider the basis of causal model information or whether they are actually sensitive to the learning data. Indeed, Bes et al. (2012) claimed that participants largely disregard data. We therefore manipulated the size of the contingency presented in Phase 1 between subjects.

Method

Participants

Two hundred ninety participants (146 female, 142 male, two participants indicated neither male nor female, $M_{Age} = 34.41$, $SD_{Age} = 11.85$) recruited via *Prolific* participated in this study and provided valid data. The chosen sample size allows us to detect medium main effects and interactions (although we did not predict any interaction effects) of $d = 0.5$ with more than 80% probability. We applied similar inclusion and exclusion

[5] There is a debate in the causal Bayes net literature about whether all causal relations are deterministic but only appear probabilistic because of unobserved causal influences (i.e., quasi-determinism), or whether causal power can be genuinely probabilistic (Cheng, 1997; Pearl, 2000; Spirtes et al., 1993). In the former case external variables are fully responsible for nondeterministic causal relations, whereas in the latter case they have the potential to modify a stable preexisting probabilistic power. For our predictions it does not matter which view is endorsed.

Table 5
Contingencies Presented to Participants in Experiment 3

Contingency	*n* (J, L)	*n* (J, ¬L)	*n* (¬J, L)	*n* (¬J, ¬L)	*P*(L|J)	*P*(L|¬J)	Δ*P*
High	18	6	5	19	0.75	0.21	0.54
Low	11	13	5	19	0.46	0.21	0.25

criteria as before. Participants were paid £0.70 for their participation.

Design, Materials, and Procedure

The study design was a 2 (causal model: direct vs. indirect [interpolated] causal relationship) × 2 (information about disablers: implicit vs. explicit) × 2 (contingency: high vs. low; see Table 5) between-subjects design. An example video illustrating the experimental procedure in one of the conditions can be accessed at https://osf.io/wqrkd/. The instructions were similar as in the previous studies. Participants were first presented with trial-by-trial learning data showing the contingency between JPH3 and Lipogastrosis. The contingencies we tested are shown Table 5. The contingency in the "high contingency" condition was the same as in Experiments 1 and 2. The "low contingency" condition had a lower predictive probability but the same probability of the effect in the absence of the cause.

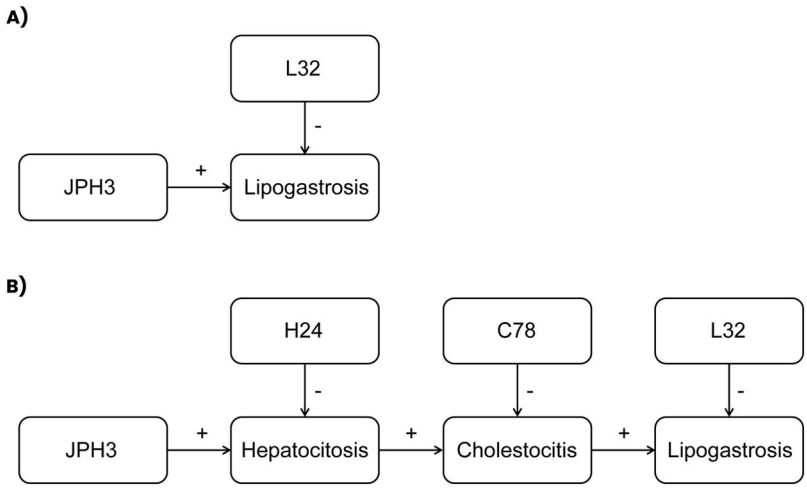
After the learning phase, participants were provided with causal model information. The direct causal condition in which no information about disablers was presented was comparable to that of Experiment 1a. In the condition in which we explicitly mentioned disablers, participants were shown the illustration depicted in Figure 10A 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 with 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.

Like in Experiment 1b, participants in the indirect (interpolated) causal condition were told that the causal model was a four-variable causal chain with the interpolated variables being Hepatocitosis and Cholestocitis. In the condition in which we explicitly mentioned disablers, participants were shown the illustration depicted in Figure 10B and read the following text:

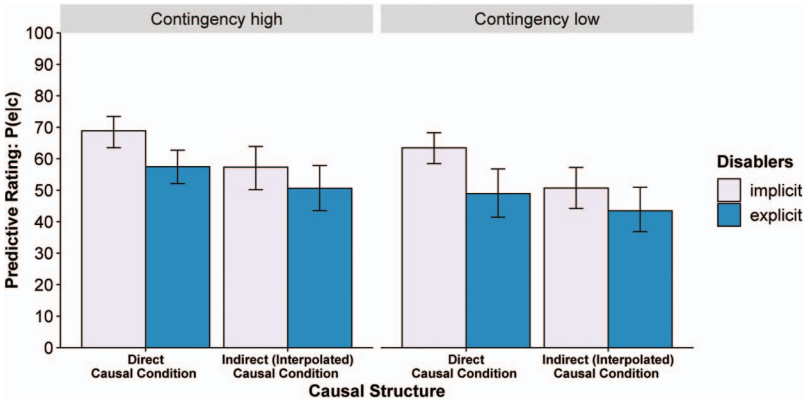
> Please read the following new information:
>
> Later, the biologists found out that JPH3 and Lipogastrosis are indirectly causally related by a chain that is illustrated in the figure below. Specifically, they found out that individuals who suffer from a JPH3 mutation have a higher probability of contracting Hepatocitosis (indicated by the arrow that goes from JPH3 to Hepatocitosis and the plus sign above the arrow), which is an abnormal increase of hepatic enzymes. Further, for individuals who suffer from Hepatocitosis there is a higher probability of contracting Cholestocitis, an abnormal

Figure 10
Illustration of the Causal Models Participants Were Shown in the Direct (A) and the Indirect (Interpolated) Causal Condition (B) After the Learning Phase in Experiment 3 (explicit conditions)



Note. In the direct causal condition one disabler was explicitly presented. In the indirect (interpolated) causal condition, three disablers were introduced.

Figure 11
Results (Means and 95% CIs) of Experiment 3



Note. See the online article for the color version of this figure.

increase in cholesterol levels. Finally, Cholestocitis 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 Hepatocitosis (indicated by the arrow with the minus sign). Having the gene C78 also reduces the probability of Lipogastrosis by reducing the probability of Cholestocitis. Finally having the gene L32 reduces the probability of Lipogastrosis.

After participants were presented with the causal model information, they proceeded to the test question. We asked the same predictive probability question as in Experiment 1a.

Results and Discussion

The results are summarized in Figure 11. A table with the exact values is provided at https://osf.iog/xq5et/. As can be seen in Figure 11, we replicated the weakening effect. The ratings for the indirect (interpolated) causal conditions were lower than for the direct causal conditions. A 2 (causal model: direct link vs. indirect (interpolated) causal relationship) × 2 (information about disablers: implicit vs. explicit) × 2 (contingency: high vs. low) between-subjects ANOVA confirmed that the effect of causal model was significant, $F(1, 282) = 15.80$, $p < .001$, $d = .46$. We also found that ratings differed depending on whether participants were informed about the existence of potential disablers or not. The predictive probability ratings were lower in the condition in which we explicitly mentioned the existence of disablers, $F(1, 282) = 18.69$, $p < .001$, $d = .51$. This finding suggests that participants in the explicit conditions exaggerated the possibility that disablers may disrupt the causal process, and therefore tended to lower their predictive probability ratings.

We also found an independent weakening effect for the indirect (interpolated) chains in the implicit conditions. In these conditions, direct causal relations also yielded higher predictive probability ratings than indirect relations (planned contrast analyses: $t(282) = 2.45$, $p = .007$ [one-tailed], $d = 0.62$, for the contingency high condition, and $t(282) = 2.90$, $p = .002$ [one-tailed], $d = 0.68$, for the contingency low condition). The pattern of findings is consistent with our hypothesis that presenting a chain of labeled variables in the indirect conditions also highlights the possibility of external influences compared with the direct causal conditions, although to a generally lesser degree than in the conditions in which disablers are explicitly mentioned.

Finally, we found an effect of contingency. Ratings were overall higher in the conditions with the higher contingency, $F(1, 282) = 9.03$, $p < .01$, $d = .35$, which shows that participants were somewhat sensitive to the data. The ANOVA yielded no significant interaction effects. The finding that learners paid attention to the data contradicts Bes et al.'s (2012) claim that participants generally ignore learning data. However, we also observed substantial deviations from the presented probabilities, some of which might be explained by conservatism effects (see Rottman & Hastie, 2014).

Experiment 4

The hypothesis that multiple variables on a chain should lead to weakening because more disablers might be considered in the indirect representation implies that participants misrepresent interpolations as additions of new variables. If participants represented the task normatively as an interpolation of preexisting mechanisms, they should adapt the causal strength estimates of the interpolated links in a way that counteracts the impact of external variables and preserves the covariation observed in Phase 1. If participants correctly represented the discovery of mechanism variables as a process in which a preexisting but previously unknown mechanism is brought to light, no weakening should occur regardless of how many links are being discovered.

Because in interpolation cases 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 (interpolated) causal conditions, both representations should normatively reflect the same impact of disablers. If participants represented interpolated chains normatively, the impact of disablers on the direct causal relation should be identical to the sum of impacts of the disablers in the interpolated chain. However, to accomplish this, participants need to understand that the causal strengths of individual links become stronger, the more links are interpolated. If participants do not understand

Table 6
Contingency Presented to Participants in Experiment 4

Observations				Conditional probabilities		Contingencies		
n (J, L)	n (J, ¬L)	n (¬J, L)	n (¬J, ¬L)	P(L|J)	P(L|¬J)	ΔP	$M\ \Delta P_{n=2}$	$M\ \Delta P_{n=4}$
10	14	0	24	0.42	0.00	0.42	0.65	0.81

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.

this crucial property of interpolations, and do not modify link strengths relative to the length of the interpolated chain accordingly, then additional interpolated links indeed increase the represented impact of disablers and hence lead to a weakening effect. Experiment 4 directly tests participants' beliefs about the strengths of links in interpolation scenarios and provides further evidence for our theory of how interpolations are misrepresented.

The goal of Experiment 4 was to test participants' belief of the causal strengths of interpolated links more directly. The key hypothesis was that people misrepresent interpolations as situations in which newly discovered variables are added to the mechanism instead of being discovered. One way to test this hypothesis is by focusing on the probabilistic relations people assume for the individual links within a causal chain. In an interpolation scenario, the covariation between the initially presented cause C and effect E should remain stable because the mechanism mediating the covariation was already operative all along prior to its discovery. When interpolating variables, the strengths of the newly introduced links should therefore go up proportional to the number of interpolated links. This is a basic consequence of the multiplication rule (Equation 1). For example, Table 6 shows that in a case in which the contingency between the initially presented two variables is 0.42 the link contingencies should on average increase to 0.65 when one variable is interpolated and to 0.81 when three variables are interpolated.

However, if people misrepresent discoveries as additions of variables to a fixed causal model, we expect to see relatively invariant strength estimates for the individual links regardless of the length of the inserted chain. The size of the link estimates may be influenced by the initially observed covariation and/or some strength prior but there is no formal requirement to adapt strength estimates to the length of the chain if variables are just added.

To test our hypothesis that participants tend to misrepresent interpolations of preexisting variables, we 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 6. In this experiment, we used a contingency which implies that JPH3 is a necessary cause of Lipogastrosis (i.e., P(L|¬J) = 0). Normatively, P(L|J) should then simply be the product of the predictive probabilities of its components. However, because participants may provide a more conservative estimate when asked about the probability of an effect in the absence of the cause, we asked them to directly estimate this quantity as well. Because we did not expect this estimate to vary with conditions, our general prediction was not affected.

We contrasted a *direct causal* condition with two *indirect (interpolated) causal* conditions (a two-links and a four-links chain). As dependent variables we asked participants to provide an estimate of the probabilistic relations of the individual links. Importantly, no data were presented about the interpolated links. As in previous experiments, the causal chains were just verbally instructed. Normatively, the averaged estimates should go up with increased length of the chains (see the dotted line with the normative predictions in Figure 12 and Table 6). If, by contrast, learners view the new variables as additions, no such trend is expected.

Method

Participants

Two hundred ten participants (124 female, 84 male, two participants indicated neither male nor female, $M_{\text{Age}} = 35.55$, $SD_{\text{Age}} = 11.29$) recruited via *Prolific* participated in this study and provided valid data. The inclusion and exclusion criteria were the same as in the previous studies. The chosen sample size allows us to detect a small interaction effect of $d = 0.25$ between test query and causal structure with more than 80% probability. A strong interaction effect is predicted by the normative values but we hypothesized that we would see no effect. To have a stricter test, we therefore tested against the possibility of a small interaction effect. Participants were paid £0.70 for their participation.

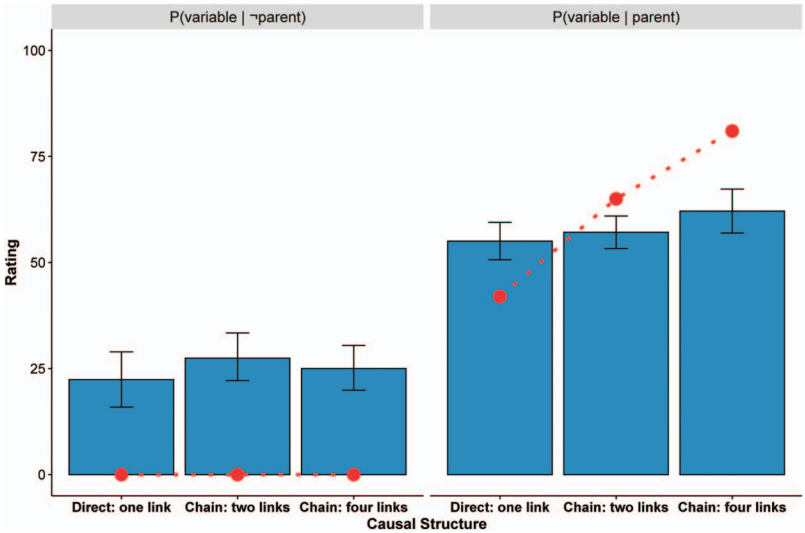
Design, Materials, and Procedure

The study had a 3 (causal model: direct vs. two-links chain vs. four-links chain; between-subjects) × 2 (type of probability query: P(variable|parent) vs. P(variable|¬parent); within-subject) mixed design. The overall procedure was comparable to those of the previous experiments. An example video illustrating the experimental procedure in one of the conditions can be viewed at https://osf.io/u9yzj/. Because the task was slightly more complex, we tried to increase attentional involvement during learning by changing the initial learning task into an active supervised learning paradigm. A separate example video showing what this active learning task looked like can be accessed at https://osf.io/qg9ka/. We first showed participants only the cause status and then asked them to actively uncover the status of the effect by clicking on a "check disease" button that was displayed on the screen. Another difference from the previously used paradigm was that participants had to navigate actively through the observations: By clicking on a "next" button they proceeded to subsequent cases while clicking on a "previous" button led them back to earlier cases. For each participant, the order in which the 48 cases were shown was randomly determined.

After the learning task, participants were given the causal model information. The information shown to participants in the direct causal and in the two-links chain conditions was like the ones we used

Figure 12

Results (Means and 95% CIs) of Experiment 4 and Normative Values Shown in the Dotted Lines



Note. See the online article for the color version of this figure.

in previous experiments. In the four-links chain condition, participants were presented with the following text and a corresponding illustration showing the described causal model:

> 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 Hepatocitosis, an abnormal occurrence of hepatic enzymes. This is indicated by the arrow (with a + sign) that goes from JPH3 to Hepatocitosis. Furthermore, Hepatocitosis can sometimes lead to Spirillicitis, an infection with bacterium colonizing the gut. This is indicated by the arrow (with a + sign) that goes from Hepatocitosis to Spirillicitis. Furthermore, Spirillicitis can sometimes lead to Cholestocitis, an abnormal occurrence of cholesterol. This is indicated by the arrow (with a + sign) that goes from Spirillicitis to Cholestocitis. Finally, Cholestocitis can sometimes lead to Lipogastrosis. This is indicated by the arrow (with a + sign) that goes from Cholestocitis to Lipogastrosis. Other factors can also influence the disease.

The order of the interpolated variables in the four-links chain (Hepatocitosis, Spirillicitis, and Cholestocitis) condition was randomized between participants. The label of the interpolated variable in the two-links chain condition was randomly varied between participants and could be any of the interpolated variable labels we used in the four-links chain condition.

After participants had been instructed about the underlying causal model, they proceeded to the test screen where we asked them to make two ratings assessing the assumed strength of a single link of the introduced causal model. Because no new data were presented, it was not possible to distinguish between links. We therefore instructed participants to assume that all links were equally strong and that they will be asked about one randomly selected link. For example, in the four-links chain condition participants 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 Lipogastrosis. Please assume that the four causal links (expressed by the four arrows) in the graph represent equally strong causal relations.

Subsequently, we asked participants to estimate the probability that a particular effect of the introduced causal model was present given that its direct parent variable was present, *P*(variable|parent), as well as the probability that this effect was present if the direct parent variable was absent, *P*(variable|¬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 Hepatocitosis has Spirillicitis?
>
> Assuming that all links are equally strong, how likely do you think is it, for example, that a mouse not having Hepatocitosis has Spirillicitis?

Ratings for both questions were provided on a slider ranging from 0 to 100 (with endpoints labeled "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 participants.

Results and Discussion

The results are summarized in Figure 12. A table with the descriptive statistics can be found at https://osf.io/dys4r/. The left panel in Figure 12 shows the results for the probability ratings in the absence of the cause, that is, *P*(variable|¬parent), whereas the right part shows the results for the predictive probability estimations, *P*(variable|parent). Given that we counterbalanced which link each participant was asked about, we here report the average of participants' estimates. The dotted line on the right side depicts the normative (average) values for the averaged predictive probabilities (*P*(variable|parent)). Participants gave overall higher ratings

when they were asked to estimate *P*(variable|parent) than when they were asked to estimate *P*(variable|¬parent), which corresponds to what they have seen. However, it can also be seen that the *P*(variable|¬parent) ratings were notably higher than the normative value (which was 0) but relatively invariant across the conditions, which entails that normatively the estimates of *P*(variable|parent) should go up with increased length of the interpolated chain. Therefore, in the following analyses we will focus only on participants' predictive probability estimates.

Figure 12 shows that there were no differences of the averaged ratings for *P*(variable|parent) between the direct and interpolated two-links chain conditions. Descriptively, there was a slight upward trend for the four-links conditions, but the estimates are much closer to the other estimates than to the normative value. A 3 (causal model: direct causal condition vs. two-links chain vs. four-links chain; between-subjects) × 2 (type of probability query: *P*(variable|parent) vs. *P*(variable|¬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*(variable|parent) ratings were higher than *P*(variable|¬parent) ratings. There was no significant interaction between "causal model" and "type of probability query," which would have been normatively expected. Post hoc tests (Scheffé tests) revealed that the *P*(variable|¬parent) ratings as well as the *P*(variable|parent) ratings did not differ from each other.

In sum, the results of Experiment 4 confirm our prediction that in interpolation scenarios people tend to assume fairly invariant link strengths, which entails a weakening effect proportional to the length of the chain. These results are consistent with those of previous experiments and support the hypothesis that people have a hard time understanding a key normative implication of causal interpolation, the predicted rise of the link strengths with an increased number of interpolated links.

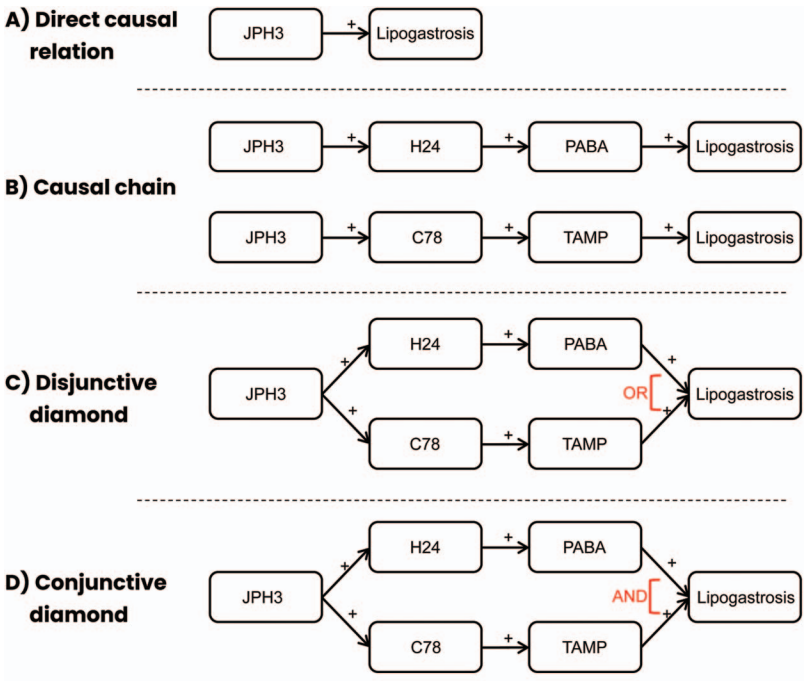
Both Experiment 3 and 4 jointly highlight the hypothesized mechanism underlying the weakening effect in interpolations. Given that participants misrepresent discoveries of preexisting variables as the addition of new variables, no adaptation of the estimated link strength (Experiment 4) is expected. Moreover, given that new variables are added to an otherwise invariant causal model, further variables potentially disturbing the causal process seem likely (Experiment 3). Both mechanisms jointly lead to the weakening effect.

Experiment 5

In the previous experiments we focused on the contrast between direct causal relations and indirect (interpolated) causal chains with various lengths. Chains are only one example for a structure linking two variables. The two variables may also be mediated by more complex causal models, for example, by causal diamond structures, which link the two variables through two parallel converging chains. Experiment 5 focuses on such more complex causal models.

For the same reasons as with simple chains, interpolations of a complex causal models should not 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 the complex causal model all along, fragments of which were only later discovered (i.e., causal stability). By contrast, if the complex causal

Figure 13
Causal Graphs Shown in the Different Conditions of Experiment 5



Note. Whether participants were presented the upper or lower causal chain in the chain condition was randomized. We did this because we wanted to present participants with the same labels and variable descriptions that were used in the two causal-diamond conditions. See the online article for the color version of this figure.

model is generated as a result of inserting a causal network between the two learning variables that was not operative prior to its discovery, then the predictive probabilities and contingencies between the initially known two variables will be dependent on the causal model mediating the two variables.

For example, consider the causal models in Figure 13A–13D, which we have used in Experiment 5. 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 or context changes). By contrast, if the causal network represented variables that did not mediate the observed covariation in Phase 1 but were added in Phase 2, thus turning the direct causal relation into a complex indirect one, then the structure of the inserted causal model (Figures 13B, C, or D), the causal strengths of the added links and their functional form is expected to alter the initially observed probabilities.

As for functional form, given that the graphs in Figures 13C and 13D convey the information that there are two chains emanating from JPH3, the open question here is how the common-effect structure converging on the effect Lipogastrosis combines the two causes, PABA and TAMP. We compared a standard disjunctive structure (disjunctive noisy-OR; Figure 13C), in which the two causes can individually as well as jointly cause the effect, with a conjunctive structure (conjunctive noisy-AND; Figure 13C), in which both causes have to be present to cause the effect. Causal links were always instructed to be probabilistic rather than deterministic.

Because Experiments 3 and 4 supported the theory that learners tend to misrepresent discoveries as changes of the underlying causal model, we expected that the ratings for the JPH3-Lipogastrosis relation will be affected by the type of the mediating causal model (i.e., a weakening effect). However, depending on whether the two chains in Conditions C and D are viewed as alternative routes to the effect, that can compensate each other (disjunctive noisy-OR) or as a conjunctive noisy-AND structure that requires both causes to be present, we predicted different sizes of the weakening effect. The lowest ratings were expected in the conjunctive noisy-AND model because here a relatively low probability of the effect should be expected.

Method

Participants

Three hundred twenty-nine participants (208 female, 121 male, M_{Age} = 35.01, SD_{Age} = 10.77) were recruited via *Prolific* and provided valid data. This sample size allows us to detect a small main effect of d = 0.40 for the causal model factor with more than 80% probability. The inclusion and exclusion criteria were the same as before. Participants were paid £0.70 for their participation.

Design, Materials, and Procedure

The type of causal model (four levels: direct causal condition vs. chain vs. conjunctive diamond vs. disjunctive diamond) was manipulated between subjects. An example video illustrating the experimental procedure in one of the conditions can be accessed at https://osf.io/wdguf/. We used the same learning paradigm as in Experiments 1 to 3. Thus again, in Phase 1 participants learned about a covariation between JPH3 and Lipogastrosis.

The learning data were the same as in Experiment 1a (see Table 1). Next, the causal model instructions were introduced. The causal model descriptions that participants were shown in the different conditions can be accessed via https://osf.io/d548h/. The graphs that were presented together with the instructions are shown in Figure 13.

Figure 13B 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 manipulated between participants. 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 participants had studied the causal model information, they proceeded to the test question, which was similar to those asked in previous experiments. Participants 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. Participants then were asked to estimate the probability that the test mouse would be suffering from Lipogastrosis.

Results and Discussion

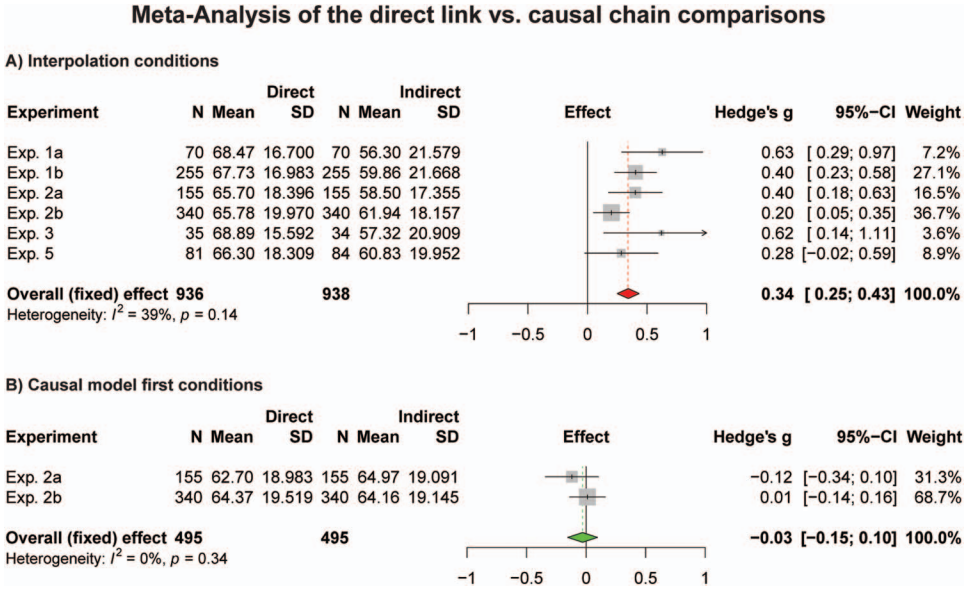
The results of Experiment 5 are summarized in Table 7. Ratings in the causal chain condition were again lower than the ratings in the direct causal condition, replicating the weakening effect. However, the difference was smaller than in most of the previous experiments (see Figure 14 in the General Discussion). The ratings in the conjunctive and disjunctive diamond conditions were also lower than the ratings in the direct causal condition, implying that the weakening effect also occurs in other interpolated causal model conditions. Moreover, as predicted, ratings differed between the conjunctive and the disjunctive diamond conditions. When participants 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

Table 7
Summary of the Results of Experiment 5

Descriptive statistics	Direct causal relation	Causal chain	Disjunctive diamond	Conjunctive diamond
M	66.30	60.83	58.82	53.00
SD	18.31	19.95	18.58	19.54
Mdn	70.00	60.00	60.00	52.00
95% CI	[62.25, 70.35]	[56.5, 65.16]	[54.79, 62.88]	[48.68, 57.32]

Figure 14

Meta-Analytic Summary of the Weakening Effects in the Different Experiments



Note. (A) Interpolation conditions. (B) Causal model-first conditions from Experiments 2a and 2b. See the online article for the color version of this figure.

two different factors independently can cause the effect (disjunctive case).

The statistical analyses confirm these descriptive patterns. First, a global one-way ANOVA was significant, $F(3, 325) = 6.69$, $p < .001$, $d = .51$, confirming that participants did indeed make different predictive probability judgments in the four different causal model conditions. Planned contrasts (one-tailed) revealed that the difference between the direct and the chain conditions was significant, $t(325) = 1.84$, $p = .034$, $d = 0.29$. If we compare the means shown in Table 7 with the ones obtained in Experiment 1, we see that the relatively small difference between the direct and the chain condition was not solely attributable to higher ratings in the chain condition but also to lower ratings in the direct causal condition.

The difference between the direct and disjunctive diamond conditions was significant, $t(325) = 2.51$, $p = .006$ (one-tailed), $d = 0.41$. The difference between the conjunctive and the disjunctive diamond conditions was also significant, $t(325) = 1.95$, $p = .026$ (one-tailed), $d = 0.31$. Ratings in the conjunctive diamond condition were significantly lower than the ratings in the direct causal condition, $t(325) = 4.43$, $p < .001$ (one-tailed), $d = 0.70$. Finally, ratings in the conjunctive diamond condition were lower than ratings in the chain condition, $t(325) = 2.63$, $p = .004$ (one-tailed), $d = 0.40$.

In sum, this experiment found that interpolations lead to weakening effects also with more complex interpolated causal models. Normatively, no difference whatsoever should have been observed between causal model conditions, had participants a full grasp of interpolations. The strong main effect of the causal model factor clearly supports the hypothesis that people do not fully understand interpolations and rather misrepresent them as augmentations of a previously simpler causal model. 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 than in the other conditions, which is reflected in lowered ratings.

General Discussion

Causal knowledge is not static; it is constantly changing based on new evidence. The present set of studies explores one important case of causal belief revision: 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 that aspirin relieves headache as a direct causal relation in daily life but later become interested in learning *how* aspirin exerts this effect.

Causal Belief Revision and the Weakening Effect

The key finding of our study is that interpolations tend to be misrepresented, which leads to the paradox of knowing more: The more we know about the mechanism, the less probable we seem to find the effect given the cause. Interpolating a mechanism should normatively not alter the statistical relationship between the variables that initially is represented as direct as long as it is assumed that the mechanism already was in place prior to its discovery. Weakening turned out to be a reliable effect that we found in all experiments. Figure 14A provides a meta-analytic summary of the effect sizes of the weakening effect. Moreover, we also list the findings obtained in Experiments 2a and 2b that weakening disappears when the causal model underlying the mechanism is instructed prior to the learning phase (Figure 14B).

The interpolation task involves belief revision based on new findings, which raises the question whether weakening is a rational response to an increase of the sample size. In the introduction we discussed several possibilities of how an extension of the original sample might alter the probability estimates. We showed that in most cases it is reasonable to assume a symmetric confidence interval that does not favor weakening over other possible changes (e.g., strengthening), which we never observed. Possible context changes were also discussed although previous research has shown that learners only tend to consider context changes when they have been saliently instructed (Cheng & Lu, 2017; Liljeholm & Cheng, 2007). But even if participants consider context changes, it was argued that they also do not favor weakening over other possible changes.

The remaining possibility that we eventually postulated as the reason for weakening is that participants may misrepresent interpolations (see Figure 5). The correct representation of interpolations 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 their causal relation. If the discovery correctly identifies variables on the mechanism path, these variables were already in place and effective prior to their discovery (see Figures 4, 5A). Or put differently: *Interpolations do not add mechanisms, they just make them known.* Thus, if the covariation between two variables has been mediated by the later discovered mechanism all along, discovering fragments of the mechanism later should not alter the covariation, unless new data have been collected that contradict the initial findings.

Our findings are consistent with the representation shown in Figure 5B, which suggests that participants start with a two variable model in Phase 1, and then, after being instructed about the discovery of mechanism variables, augment the initial causal model by adding the discovered variables to an otherwise invariant causal model. In Experiments 3 and 4 we tested implications of this theory. The theory predicts that the addition of a variable may sensitize participants to possible additional external influences that potentially disrupt the causal process, for example disablers. This is demonstrated in Experiment 3.

Experiment 4 focuses on the difference between adding a new variable versus discovering a preexisting variable. If variables are added, it is reasonable to assume that the causal strengths of the links are not affected by the number of added variables, which should lead to a weakening proportional to the number of added variables with probabilistic links. By contrast, increasing the length of a discovered preexisting mechanism (Figure 5A) should lead to adaptations of the causal strength estimates regarding the newly discovered links in the mechanism to preserve the overall covariation of the variables observed in Phase 1. The more mechanism variables are discovered, the stronger the causal strengths of the new mechanism links should on average become. If causal strength estimates are correctly adapted, the additional disablers associated with the interpolated variables should not weaken the causal relation of the two target variables because increasing causal strength implies that the impact of each mediating disabler will be diminished.

We also considered a heuristic model that can be derived from a modular approach to Bayes nets. On a modular approach, people should view the causal links as independent and use causal strength priors when extending models. This approach predicts that people use the heuristic that indirect relations tend to be weaker than direct ones. Unlike our belief revision account, this heuristic does not distinguish between interpolations of preexisting mechanisms and additions of new variables and is insensitive to the order in which direct and indirect relations are presented. We refuted this heuristic as an account of our task in Experiment 2, in which we manipulated assumptions about the underlying causal models (direct vs. indirect) between subjects. We showed in this experiment that the weakening effect can only be seen in the interpolation task in which learners acquired knowledge about the covariation between two variables prior to being informed about the underlying causal model (direct vs. indirect). When the order was reversed, no weakening effect was observed. Thus, we did not see evidence for the claim that in a between-subjects manipulation participants generally view indirect relations as weaker than direct ones given identical learning data.

We have used causal Bayes net theory as a framework for our analysis of interpolations and causal discoveries. This framework conveniently combines structure and strength information and is therefore useful for making predictions about probability estimates based on causal knowledge. However, our general findings are not tied to the assumption that learners literally use causal Bayes nets when reasoning about mechanisms. In the past decades, alternative theories have been proposed, sometimes in opposition to causal Bayes nets. For example, recent mechanism theories do not model causation as dependency relations between events; instead, they start with an analysis of entities and their activities which are organized such that they are productive of regular changes (e.g., Machamer et al., 2000). Such a theory might, for example, describe the mechanism of chemical neurotransmission as beginning with a presynaptic neuron that transmits a signal to a postsynaptic neuron by releasing neurotransmitter molecules. Mechanism accounts start with the analysis of the structure of the hardware and the capacities of its components, which jointly give rise to causal regularities (see also Cartwright, 1999; Glennan, 2017).

Generally, our predictions could also be cast within the framework of these mechanism theories. Given that learners in our task learn *probabilistic* relations between causes and effects, it would be necessary to include the assumption that we often have only partial knowledge of mechanisms and that unobserved factors may influence the behavior of observed activities, either externally or as mediating components. Given such an augmented representation of mechanisms, it is easy to translate our hypothesis in this framework. Normatively interpolations are then discoveries of mediating mechanism components that had not been known before (such as additional components that mediate neural signaling). The hypothesis predicting the weakening effect could be explained by the assumption that learners may have a tendency to represent discoveries of preexisting mechanisms as the addition of new mechanism variables.

In sum, our predictions are largely independent of the way causal processes are represented. Different theories can be adapted to derive the same qualitative predictions. Specific constraints of causal Bayes nets, such as the Markov condition, are not necessary to analyze interpolations and the weakening effect, although causal Bayes nets have the advantage at the moment that they explicitly link causal structures with probabilities and therefore allow for more precise predictions.

The Weakening Bias and Causal Stability

The majority of studies on biases in human reasoning compare judgments with how normative accounts, such as probability theory or logic, would represent the tasks at hand (see Newell et al., 2015, for an overview). In some cases, multiple normative accounts compete so that one goal for the psychologist is to coordinate participants' understanding of the task with the most adequate normative account (see Skovgaard-Olsen et al., 2019, for an analysis of this process). In contrast to cases in which probability theory or logic are taken as the normative standard, the bias we discovered is based on an assessment of whether observed patterns of belief revision are consistent with objective causal features of the world, which is a topic of philosophical theories of causation and scientific theories. We have discovered a mismatch here. The weakening effect suggests that participants misrepresent interpolations as additions of variables and not as discoveries of preexisting mechanisms. Thus, participants seem to routinely violate the causal stability assumption. When scientists discovered the role of prostaglandins in relieving headaches, they believed that prostaglandins already played a role prior to their discovery and did not show up all of a sudden, which would make the previously observed covariation seem magical.

It is important to note that the causal stability assumption is not derived from axiomatic accounts but is an assumption about the world. If we lived in a world in which discoveries would insert mechanisms into preexisting simpler causal models, the weakening effect would not be a bias but a correct reflection of how the world operates. We believe that most researchers in natural kind domains, such as biology, physics, physiology, or chemistry, share the stability assumption we have postulated.

However, there is one alternative position coming from a representative of social constructionism (Latour, 2000).[6] Bruno Latour discusses the case of Ramses II, the third pharaoh of the Nineteenth Dynasty of Egypt. In 1975 the pharaoh's mummy was medically examined by French doctors in Paris. Among other diseases, they discovered signs of tuberculosis, which, as Robert Koch has discovered in 1882, is caused by a bacillus. The question Latour asks is whether Ramses II had tuberculosis and whether it was caused by the bacillus that has been discovered three thousand years after his death. Latour questions this when he writes: "The attribution of tuberculosis and Koch's bacillus to Ramses II should strike us as an anachronism of the same caliber as if we had diagnosed his death as having been caused by a Marxist upheaval, or a machine gun, or a Wall Street crash" (p. 248). We believe that both scientists and nonscientists would disagree with Latour, who treated natural kinds and artifacts (e.g., machine guns) as analogous. Unlike the Koch bacillus which plausibly already existed when Ramses II contracted tuberculosis, machine guns were invented in the 19th century and were therefore not part of the world of Ramses II. Given the medical examination in Paris, the most plausible assumption is indeed that Ramses II suffered from tuberculosis and that the disease was caused by a bacillus. This backward inference seemed to be triggered by our belief in causal stability. It is an interesting question whether weakening would be a rational inference for a social constructionist, such as Latour.

Although our research is the first that discovered a bias in the context of causal discovery and the representation of interpolation, other researchers have studied with other tasks whether probability inferences made in the context of causal reasoning conform to normative accounts or are biased (see Waldmann, in press, for a review). The majority of studies concludes that most deviations can be explained as rational if the right assumptions about background knowledge participants bring to bear on the task are made. However, unlike in the case of interpolations, the results are often ambiguous.

For example, research on reasoning with causal conditionals has also studied belief revision (Oaksford & Chater, 2013, 2017). One central theory of reasoning with conditionals subscribes to what has been called "The Equation" (Edgington, 1995). According to this theory, the probability of a conditional is equated with a conditional probability, $P(\textit{if } p \textit{ then } q) = P(q|p)$. According to the Ramsey test, one supposes that p = True, adds it to one's belief basis, makes revisions, and then reads off the probability of q (see Evans et al., 2003; Over et al., 2007). Using Bayesian conditionalization, one can then update the belief in q when learning that p is true by using the conditional probability $P(q|p)$. For example, if we assume as conditional premise that "if the key is turned, then the car starts," we can use the associated conditional probability to make a probabilistic inference from "the key is turned" to "the car starts." Bayesian conditionalization assumes invariance, which is related to our causal stability assumption (Jeffrey, 1992). The invariance assumption assumes that the conditional probabilities stay the same between the old and new distributions, that is, $P_1(q|p) = P_0(q|p)$. A number of empirical studies have shown that people do not always conform to this assumption (Oaksford & Chater, 2017; Zhao et al., 2012; Zhao & Osherson, 2010). When confronted with cases that appear to be counterexamples of the conditional (e.g., a car that does not start despite the key being turned), thus suggesting the operation of disablers, reasoners often seem to revise the conditional probability estimates to a lower value. At first sight, this seems to be a case of weakening.

However, whether weakening is rational or irrational in this case can be debated. On the one hand, one could argue that the inferences may be biased because reasoners do not take into account that counterexamples are normative implications of probabilistic relations. If smoking probabilistically causes cancer in a large sample of cases, a single case of a healthy smoker should be viewed as a natural consequence of the probabilistic relation, not as a counterexample that should lead to a major revision of the probability. It may well be that the finding that people revise too fast may be a sign of a faulty understanding of probabilities. However, the alternative, defended by Oaksford and Chater (2017), is equally plausible. Participants may infer that the car that does not start is a counterexample that belongs to a different population, cars that are broken. Using background knowledge, people might then validly infer that the probabilities that have been learned in the context of intact cars are invalid in this different population because disablers are at work that in the regular population are not present.

We believe that the bias we discovered in the context of interpolations cannot be similarly explained away by background knowledge. One key difference is that the findings in the research about reasoning with conditionals presented participants with cases in which causes and effects were both instructed as either

[6] We thank Ralph Hertwig for pointing us to this article.

present or absent. These patterns might have led participants to reason about possible factors explaining the individual case. Moreover, it is possible to represent each case as a singular case or as being part of a separate subcategory. By contrast, in our paradigm we instructed participants that they should envision a *randomly sampled* test case in which the cause is present and then estimate the probability of the effect. The status of the effect was left open. Thus, our task is more similar to the request to assess the likelihood of cancer given smoking than to asking about smokers and nonsmokers with and without cancers, which may trigger different explanations based on background knowledge. Moreover, we have shown that background knowledge about disablers should in our task only lead to weakening if participants confuse discoveries of preexisting mechanisms with additions of variables. Had participants correctly understood what interpolations entail for causal strength and for disablers, they should have inferred that the observed covariation was mediated by the mechanism all along. In our view, the weakening effect is therefore one of the first findings in causal reasoning research that avoids the ambiguity of previous research and demonstrates a clear case of biased reasoning.

Perspectives for Future Research

A key question that deserves further research concerns the psychological mechanism underlying the weakening effect. We argued that participants' inferences were consistent with the hypothesis that they treated causal discovery as a situation in which further variables were added to an otherwise invariant causal model, which violates causal stability assumptions. Although it is possible that participants actually believed that discoveries add variables as some social constructionists do (Latour, 2000), our experience when presenting the project is that most members of the audience agree that interpolations in natural kind domains target preexisting mechanisms. Indirect evidence for this was presented in Experiment 2, which showed that participants had a perfect understanding of how a mechanism is connected to the two learning variables when it is presented prior to the learning phase (cf. Figure 14B), whereas we observed a weakening effect when causal model information was presented subsequently (cf. Figure 14A). This pattern of results demonstrates that the sequence of learning is crucial in this task. Therefore, a more plausible explanation of the weakening effect is that participants have difficulties with retroactively inferring the operation of a mechanism that had been unknown until its discovery. To realize that a discovered mechanism had already been present in the past requires to retrospectively insert that mechanism in the appropriate position in the causal structure, as well as to make the right adjustments concerning causal strength and the impact of surrounding variables (e.g., disablers). This is effortful and probably, owing to an attempt to reduce effort, there seems to be a tendency to confuse lack of knowledge about a variable with the assumption of its absence (see Waldmann, 2000; Waldmann et al., 2012). We proposed a simulation mechanism that explains the findings. If participants run mental simulations to incorporate the new information provided in Phase 2, it is easier to add a new variable while leaving the model used in Phase 1 intact than to retroactively resimulate the Phase 1 trials. Future research will have to test this hypothesis more directly.

We already discussed that a simple alternative to our belief revision theory of weakening in interpolation tasks is a heuristic which asserts that people may have the general bias that indirect relations tend to be weaker than direct ones. This heuristic is plausible when the direct relations are subcomponents of indirect ones (as in Bes et al., 2012) but is inappropriate when there is no reason to make this assumption (see the introduction for examples). Experiment 2 has already shown that participants do not generally use this heuristic. It may be interesting to run further tests of the heuristic account, however. A plausible mechanism underlying the weakening effect on this account might be that people use causal strength priors both when learning about the covariation in Phase 1 and when making assumptions about the strengths of the interpolated links in Phase 2. Such an account entails a weakening effect if it is assumed that uncertainty about the covariation in Phase 1 allows the priors of the interpolated causal links to revise the initial covariation estimates. An interesting prediction of this account is that with an increase of the sample size in Phase 1 uncertainty should be reduced, thus leaving less space for a downward revision of the initial estimate.[7] With increased sample size in Phase 1, thus decreasing uncertainty about the covariation observed in Phase 1, the weakening effect should become smaller. No such effect is predicted by our belief revision account, which makes an experiment manipulating sample size an interesting test.

Our main goal in the present study was on comparing direct with indirect (interpolated) causal relations. Experiment 5 adds to these studies a first examination of more complex causal models. An additional interesting question concerns partial knowledge about complex causal models. 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 participants become aware that their mechanism knowledge is partial and fragmentary.

Our focus was on natural kind domains, which we presented with unfamiliar variables. We uniformly found weakening effects but there may be tasks in which strengthening or no systematic change may be observed. One possibility to find strengthening that was already mentioned in the introduction is to systematically study other domains. Artifacts seem to be a case in which mechanism change is likely, which in this domain may lead to strengthening. Engineering generally strives to improve the reliability of artifacts rather than weaken it. For other domains, such as intuitive psychology, economics or sociology, we have few clear intuitions, which would make studying these domains interesting for future research.

In discussions with colleagues and among ourselves, 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 suggesting another way of reversing the weakening effect.[8] For example, understanding how a vaccine helps us may make its efficacy 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 can affect our probability

[7] We thank Bob Rehder for suggesting this study.

[8] We thank Aaron Blaisdell and Keith Holyoak for mentioning this possibility.

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 et al., 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 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 temper our urge to gain more knowledge. Recent research has shown that people tend to know very little about mechanisms (Chater, 2018; 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, be reluctant to find out more about mechanisms. Causal knowledge revision is certainly an important topic for reasoning research and should attract more interest in the future.

Context Paragraph

The article is a result of a long-standing collaboration between Michael Waldmann and Katya Tentori, which aims at combining expertise about causal and probabilistic reasoning. Michael Waldmann has studied causal learning and reasoning and has been particularly interested in investigating the role of causal models in learning and reasoning. Katya Tentori has studied probabilistic and inductive reasoning by combining normative analyses with psychological theories. The collaborative project started with studies about differences in the representation of causal chains versus direct causal relations. But after Simon Stephan and Stefania Pighin had joined the project, our interest shifted to a rarely investigated task, causal belief revision and the interpolation of mechanism information. We found weakening effects in both the studies about causal chains and in the new interpolation paradigm, and became interested in the question whether this effect is based on similar psychological representations. Moreover, given that not only the participants in the experiments but also many colleagues found weakening intuitive, we also wondered whether the effect can be defended as rational or whether it represents a new type of bias in causal reasoning.

References

Ahn, W., & Dennis, M. (2000). Induction of causal chains. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 19–24). Erlbaum.

Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 153–168. https://doi.org/10.1037/a0013764

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, É. (2012). Non-Bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*(7), 1178–1203. https://doi.org/10.1111/j.1551-6709.2012.01262.x

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338. https://doi.org/10.1037/rev0000061

Byrne, R. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61–83. https://doi.org/10.1016/0010-0277(89)90018-8

Cartwright, N. (1999). *The dappled world. A study of the boundaries of science*. Cambridge University Press. https://doi.org/10.1017/CBO9781139167093

Chater, N. (2018). *The mind is flat. The illusion of mental depth and the improvised mind*. Allen Lane.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367

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). Oxford University Press.

Crupi, V., Tentori, K., & Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, *116*(4), 971–985. https://doi.org/10.1037/a0017050

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23*(5), 646–658. https://doi.org/10.3758/BF03197265

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*(3), 274–282. https://doi.org/10.3758/BF03211151

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329. https://doi.org/10.1093/mind/104.414.235

Evans, J. St. B. T., Handley, S. H., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 321–335. https://doi.org/10.1037/0278-7393.29.2.321

Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, *31*(5), 800–815. https://doi.org/10.3758/BF03196118

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press. https://doi.org/10.1093/oso/9780198779711.001.0001

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. https://doi.org/10.1037/a0017201

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. https://doi.org/10.1177/0963721412447619

Jara, E., Vila, J., & Maldonado, A. (2006). Second-order conditioning of human causal learning. *Learning and Motivation*, *37*(3), 230–246. https://doi.org/10.1016/j.lmot.2005.12.001

Jeffrey, R. (1992). *Probability and the art of judgment*. Cambridge University Press. https://doi.org/10.1017/CBO9781139172394

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, *107*(2), 397–402. https://doi.org/10.1037/0033-295X.107.2.397

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. The MIT Press.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195176803.003.0011

Latour, B. (2000). On the partial existence of existing and non-existing objects. In L. Daston (Ed.), *Biographies of scientific objects* (pp. 247–269). University of Chicago Press.

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science*, *18*(11), 1014–1021. https://doi.org/10.1111/j.1467-9280.2007.02017.x

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory & Cognition*, *33*(8), 1388–1398. https://doi.org/10.3758/BF03193371

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984. https://doi.org/10.1037/a0013256

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, *39*(1), 65–95. https://doi.org/10.1111/cogs.12132

Mayrhofer, R., & Waldmann, M. R. (2016). Sufficiency and necessity assumptions in causal structure induction. *Cognitive Science*, *40*(8), 2137–2150. https://doi.org/10.1111/cogs.12318

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2015). *Straight choices*. Psychology Press. https://doi.org/10.4324/9781315727080

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608

Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*(3–4), 346–379. https://doi.org/10.1080/13546783.2013.808163

Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. In M. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 327–346). Oxford University Press.

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97. https://doi.org/10.1016/j.cogpsych.2006.05.002

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579–595. https://doi.org/10.1214/14-STS486

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

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). Oxford University Press.

Perales, J. C., Shanks, D. R., & Lagnado, D. (2010). Causal representation and behavior: The integration of mechanism and covariation. *The Open Psychology Journal*, *3*(1), 174–183. https://doi.org/10.2174/1874350101003010174

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition*, *45*(2), 245–260. https://doi.org/10.3758/s13421-016-0662-3

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). Oxford University Press.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109–139. https://doi.org/10.1037/a0031903

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflict and conditionals. *Psychological Review*, *126*(5), 611–633. https://doi.org/10.1037/rev0000150

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. Riverhead Books.

Spirtes, P., Glymour, C., & Scheines, P. (1993). *Causation, prediction, and search*. Springer-Verlag. https://doi.org/10.1007/978-1-4612-2748-9

Spohn, W. (2012). *The laws of belief: Ranking theory and its philosophical applications*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199697502.001.0001

Stephan, S., Tentori, K., Pighin, S., & Waldmann, M. R. (2020, June 15). *Interpolating causal mechanisms: The paradox of knowing more*. Retrieved from osf.io/aqzps

Taylor, E. G., & Ahn, W. (2012). Causal imprinting in causal structure learning. *Cognitive Psychology*, *65*(3), 381–413. https://doi.org/10.1016/j.cogpsych.2012.07.001

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567. https://doi.org/10.1037/0033-295X.101.4.547

von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory & Cognition*, *44*(3), 469–487. https://doi.org/10.3758/s13421-015-0568-5

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). Cognitive Science Society.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53–76. https://doi.org/10.1037/0278-7393.26.1.53

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review*, *8*(3), 600–608. https://doi.org/10.3758/BF03196196

Waldmann, M. R. (Ed.), (2017). *The Oxford handbook of causal reasoning*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199399550.001.0001

Waldmann, M. R. (in press). The rationality of everyday causal cognition. In M. Knauff & W. Spohn (Eds.), *The handbook of rationality*. MIT Press.

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 733–752). Oxford University Press.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222–236. https://doi.org/10.1037/0096-3445.121.2.222

Waldmann, M. R., Schmid, M., Wong, J., & Blaisdell, A. P. (2012). Rats distinguish between absence of events and lack of evidence in contingency learning. *Animal Cognition*, *15*(5), 979–990. https://doi.org/10.1007/s10071-012-0524-8

Waldmann, M. R., & Walker, J. M. (2005). Competence and performance in causal learning. *Learning & Behavior*, *33*(2), 211–229. https://doi.org/10.3758/BF03196064

Woodward, J. (2011). Mechanisms revisited. *Synthese*, *183*(3), 409–427. https://doi.org/10.1007/s11229-011-9870-3

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29. https://doi.org/10.1016/j.cogpsych.2014.11.001

Zhao, J., Crupi, V., Tentori, K., Fitelson, B., & Osherson, D. (2012). Updating: Learning versus supposing. *Cognition*, *124*(3), 373–378. https://doi.org/10.1016/j.cognition.2012.05.001

Zhao, J., & Osherson, D. (2010). Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking & Reasoning*, *16*(4), 288–307. https://doi.org/10.1080/13546783.2010.521695

Received June 15, 2019
Revision received October 14, 2020
Accepted October 29, 2020 ■