Explanations in Causal Chains: Selecting Distal Causes Requires Exportable Mechanisms

Jonas Nagel (jnagel1@uni-goettingen.de)
Simon Stephan (sstepha1@uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

When A causes B and B causes C, under what conditions is A a good explanation for the occurrence of C? We propose that distal causes are only perceived to be explanatory if the causal mechanism is insensitive to inessential variations of boundary conditions. In two experiments, subjects first observed deterministic A → B → C relationships in a single exemplar of an unknown kind. They judged A to be crucial for C by default. However, when they subsequently learned that the causal mechanism fails to generate the A → C dependency in other exemplars of the same kind, subjects devalued A as a crucial explanation for C even within the first exemplar. We relate these findings to the idea that good explanations pick out portable dependency relations, and that sensitive causes fail to meet this requirement.

Keywords: explanation; causal mechanisms; causal chains; sensitivity; portability

Introduction

Causal relationships are implemented by causal mechanisms (Machamer, Darden, & Craver, 2000). If we say that one event (A) causes another event (C), we generally assume that there is some process leading from A to C that can principally be discovered, even if we do not yet know what the actual mechanism is. Accordingly, the causal arrow in notations such as A → C is sometimes interpreted as a "mechanism placeholder" (Pearl, 2000).

When mechanism knowledge about a particular causal relationship (e.g., A → C) is made explicit, the resulting causal model takes the form of a causal chain (e.g., A → B → C, where B is an intermediate cause implementing the mechanism). The current research asks how such integration of mechanism knowledge affects people's conceptualization of the original causal relationship. More specifically, the question is which properties of mechanism B can affect people's impression of the importance of A for explaining C.

Intuitively, there are two ways to interpret the causal role of B in causal chains. First, B could be seen as a *mediator* implementing the causal influence of A on C. Under this reading, A continues to be seen as explanatory for C, even though its causal influence is completely mediated via B. Second, it could be seen as an *alternative explanation* for the occurrence of C, screening off the influence of A on C. The fact that, if we know the value of B, A adds nothing to explaining C, provides a reason to devalue A as appropriate explanation of C under this interpretation.

Nagel and Stephan (2015) showed that both interpretations can arise when subjects learn that different mechanisms implement one and the same type-level causal relationship. They had their subjects learn a strong dependency of grades in a gym class (C) on the pupils' gender (A) from fictional covariation data. Subjects indicated strong agreement with the claim that, within the observed sample, a pupil's gender was crucial for his or her grade. In a second learning phase, half of the subjects learned that the A → C relationship was mediated by a genetically determined physiological process (B_1), while the other half learned that it was mediated by the gender preferences of the teacher (B_2). Afterwards, subjects in the physiological mechanism condition continued to endorse the statement that a pupil's gender was crucial for this pupil's grade (indicating that B_1 was interpreted as a mediator of the original A → C relationship), while subjects in the teacher mechanism condition devalued gender as crucial explanation for the grades (indicating that B_2 was interpreted as an alternative explanation for C).

Both conditions were equivalent in terms of objective causal structure and observed dependency patterns, so it seems the difference in interpretation results from some aspect of the manifold content-related differences between both contrasted mechanisms. One salient hypothesis is that intentional agents, like the teacher, might be seen as initiators of causal sequences (unmoved movers) and therefore always screen off the influence of upstream physical preconditions of their actions from downstream effects. Blind physical processes like genetics, by contrast, may not have this quality and may thus be regarded as mere mediators of the influence of upstream root causes. In a second experiment, Nagel and Stephan (2015) ruled out this hypothesis. They presented their subjects with a scenario in which a physical signal (A) was picked up by a human agent who deliberately reacted to the signal (B) to produce a final outcome (C). Subjects repeatedly observed this deterministic causal chain in all conditions. Half of the participants learned that the agent had a benign motivation in implementing the A → C chain, while the other half learned that he had a malevolent motivation. Afterwards, they were asked to what extent it was appropriate to state that the original signal (A) vs. the agent's reaction (B) was crucial for the occurrence of the outcome (C). It turned out that in case of the benign agent, both the distal signal and the proximal action were judged to be equally crucial for the occurrence of C. The distal physical cause thus retained its explanatory power and was seen to bring about the outcome *by means of* human agency. By contrast, if the agent had a morally dubious motive, the distal cause was devalued as explanation despite a perfect dependency relation with the outcome in the observed sample; proximal human agency served as *alternative explanation* for the outcome instead. This finding strongly suggests that it is not an agent's intentionality per se that leads to devaluation of upstream causes as explanations for downstream effects, but rather some other property that is related to the motivation of the agent. In the remainder of this paper, we will outline and test the hypothesis that this property does not reside exclusively in moral qualities of intentional agents, but quite generally reflects inferences about whether the mechanism can be expected to generalize to other, inessentially different contexts.

Sensitivity and Explanatory Relevance

Woodward (2006) investigated the human practice of making causal claims. He noted that causal claims require not only that the effect be counterfactually dependent on the cause, but also that this counterfactual dependence continue to hold under varied boundary conditions. Causal relationships that do not fulfill this second requirement are called *sensitive*, and Woodward (2006) argues that sensitive causal relationships are regarded as deficient despite a strong dependence relationship under the conditions in which they do obtain. Good causes are those that not only bring about their effects in the narrow context of actually observed circumstances, but would continue to do so in different contexts. The requirement for good causes to be insensitive resonates with philosophical and psychological accounts of explanation. Many theorists have argued that good explanations tend to pick out factors that are generalizable beyond the concrete set of observations that presently is to be explained (Garfinkel, 1981; Hitchcock, 2012; Lombrozo & Carey, 2006). This ensures that the generated explanation will be useful for making predictions and for planning interventions in future similar situations (Lombrozo, 2010).

One reason for high sensitivity of causal relationships is that the mediating mechanism works reliably only under quite specific boundary conditions, but is easily disturbed in other, similar situations. We propose that whenever people find out that an observed causal relationship is implemented by a mechanism that is highly sensitive in this sense, they devalue the distal cause as explanation for the terminal effect. To illustrate, consider again the scenarios used by Nagel and Stephan (2015). If the influence of gender on grades is mediated by a genetically determined physiological mechanism, this implies that the relationship will continue to hold in future observations with different samples of pupils, which makes the A → C relationship insensitive. The teacher mechanism, by contrast, implies that this relationship depends on the presence of highly peculiar boundary conditions which will rarely be met in other, similar situations (as most other teachers, hopefully, will not exhibit the same bias). This makes the observed A → C relationship highly sensitive—it will break down as soon as we leave the narrow context of the class that was actually observed in the sample. It is recognized that gender will not *generally* influence grades and is therefore considered a poor explanation for the grades *even within the observed sample*.

The mechanisms compared by Nagel and Stephan (2015) differed on many dimensions other than sensitivity, including intentional agency and moral abnormality. Our goal in the present experiments is to isolate the sensitivity of the mediating mechanism. We created new experimental material from the domains of biology and physics in which we manipulated a given mechanism's sensitivity purely in statistical terms. We first presented subjects with a single exemplar of an unknown natural kind or artifact and let them discover a deterministic A → B → C chain within this entity. In a subsequent learning phase, we showed subjects the same exemplar again, but this time in the company of several other exemplars of the same kind with identical appearance. One half of the subjects saw that the new exemplars behaved just like the first exemplar in terms of the A → B → C dependency pattern. The other half saw that only the first exemplar once again showed the same dependency pattern, while in all other exemplars the presence of A failed to lead to the presence of B (and, hence, C). Subjects in both conditions were then asked how appropriate it was to say that A was crucial for the presence of C *within the first exemplar only* which had constantly displayed perfect dependency relations in both conditions. We predicted reduced appropriateness ratings in the condition in which the dependence of C on A did not generalize to other exemplars of the same kind.

In the first experiment, we tested and confirmed these predictions in the domain of biology. In the second experiment, we replicated the findings in the domain of artifacts and additionally controlled for the relative complexity of the potentially explanatory variables A and B.

Experiment 1

Participants

The experiment was conducted as an online study. A total of 150 subjects were recruited from a panel (*www.pureprofile.com*), 44 of which (29%) were removed prior to the analyses because they did not complete the survey or failed to solve a simple attention check question at the end. The mean age of all included subjects ($N = 106$, 80 women) was 37 years ($SD = 8.16$). Included subjects received a payment of £ 6 per hour.

Design, Materials, and Procedure

Subjects were randomly allocated to the conditions (Sensitivity: Sensitive vs. Insensitive). They were asked to take the perspective of a marine biologist who discovered a single exemplar of deep sea fish and called it "Fish #1". Their task was to test whether noise (A) leads to activity in the brain of the fish (B) and finally to an illumination of the fish's antenna (C). We then presented our subjects an animation of Fish #1 (see Figure 1). When participants pressed the "Play" button, they saw sound waves coming out of the speaker. About half a second later, the brain activity device's monitor displayed a flickering amplitude moving across the screen. Finally, about one second later, the fish's flash bulb turned from blue to yellow. If participants hit the stop button, all variables returned to their initial state. Subjects could (de-)activate the loudspeaker as often as they wished.

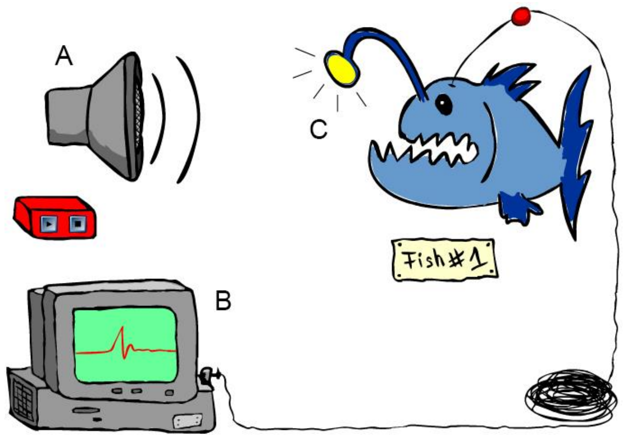

Figure 1: Screenshot of the animation used in the first learning phase of Experiment 1 while all three variables are active. Letters A, B, and C were not shown to participants.

When participants felt they had learned enough about the relationships, they proceeded to the first question screen on which they were asked how appropriate the following statements were for describing the observations they had just made of Fish #1. The first statement was "The presence of sound waves is crucial for Fish #1's antenna to lighten up" (appropriateness rating *$[A \rightarrow C]_{pre}$*), and the second statement was "Activity of the brain area is crucial for Fish #1's antenna to lighten up" (appropriateness rating *$[B \rightarrow C]_{pre}$*). We used two independent rating scales to allow participants to judge both causes as equally explanatory, while at the same inviting them to see both statements as contrastive alternatives by using the word "crucial". Participants provided their judgments on 11 point rating scales ranging from "0 = not at all appropriate" to "10 = very appropriate" for each statement. We expected equally high ratings for both statements, indicating that brain activity is seen as a mediator and sound waves are seen to be explanatorily relevant.

The experimental manipulation was applied after subjects had given their baseline ratings. All participants read that they had caught nine additional exemplars of the same kind of fish. On the next screen, they saw a large animation showing all ten fish (consecutively labelled Fish #1 to Fish #10), each in the same set-up as shown in Figure 1. Below the ten fish, there was a device with a play- and stop-button which they could use to (de-)activate all ten loudspeakers simultaneously. In both conditions, Fish #1 again reacted exactly as in the first learning phase. The crucial difference between both conditions was the behavior of the additional nine fish exemplars. In the insensitive condition, all other fish behaved exactly like Fish #1, while in the sensitive condition, none of the other fish reacted to the activation of the loudspeaker with brain activity or antenna lightning. This manipulation was intended to confirm in the insensitive condition the assumed prior expectation that the A → C dependence is exportable from Fish #1 to the whole kind, while subjects in the sensitive condition should conclude that the A → C dependence is not exportable beyond the narrow context of Fish #1.

After having made these additional observations, subjects were asked to reconsider the results of their previous experiment *with Fish #1 only* and to answer the same two questions again in light of their new knowledge about the whole swarm of fish. The appropriateness ratings *$(A \rightarrow C)_{post}$* and *$(B \rightarrow C)_{post}$* were measured exactly as the baseline ratings described above. Our central prediction for the sensitive condition was that the $(\mathrm{A} \rightarrow \mathrm{C})_{\mathrm{post}}$ ratings should drop considerably because the A → C dependency is not exportable beyond the context of Fish #1. The $(\mathrm{B} \rightarrow \mathrm{C})_{\mathrm{post}}$ appropriateness ratings, by contrast, should not be reduced by the swarm information. Brain activity remains a good predictor of antenna flashing across the whole swarm. In the insensitive condition, of course, neither of the ratings should be affected by the swarm data.

Finally, we wanted to make sure that participants encoded the contingencies between the variables accurately both within Fish #1 and across the whole swarm of fish. On a first screen, participants were prompted to recall what they had learned about Fish #1 and were asked six conditional probability questions concerning Fish #1 only. For example, $P(C|A)$ was assessed with the question "How likely is it for Fish #1's antenna to lighten up given that sound waves are present?" and an 11 points rating scale ranging from 0 (impossible) to 100 (certain). Analogous questions were asked for $P(C|\neg A)$, $P(B|A)$, $P(B|\neg A)$, $P(C|B)$, and $P(C|\neg B)$. On a second screen, the same six questions were repeated with the whole swarm as reference class. These twelve estimates were used to compute contingency estimates (ΔP) for each of the three causal relationships both at the exemplar level and at the swarm level. Equally high contingency estimates at the exemplar level for A → C and B → C would demonstrate that the expected effects on the appropriateness ratings would not be due to different dependency assumptions within the exemplar, but rather due to sensitivity of the A → C relationship across the whole kind.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 2a. We conducted a three-way 2 (Sensitivity: Sensitive vs. Insensitive, between-subjects) × 2 (Relationship: A → C vs. B → C, within-subject) × 2 (Rating Position: Pre vs. Post, within-subject) mixed ANOVA. We obtained a significant two-way Sensitivity × Position interaction, $F_{1,104} = 8.61$, $p < .01$, $\eta_g^2 = .014$, indicating that the swarm information affected appropriateness ratings in the sensitive condition more than in the insensitive condition. More importantly, this interaction was qualified by a

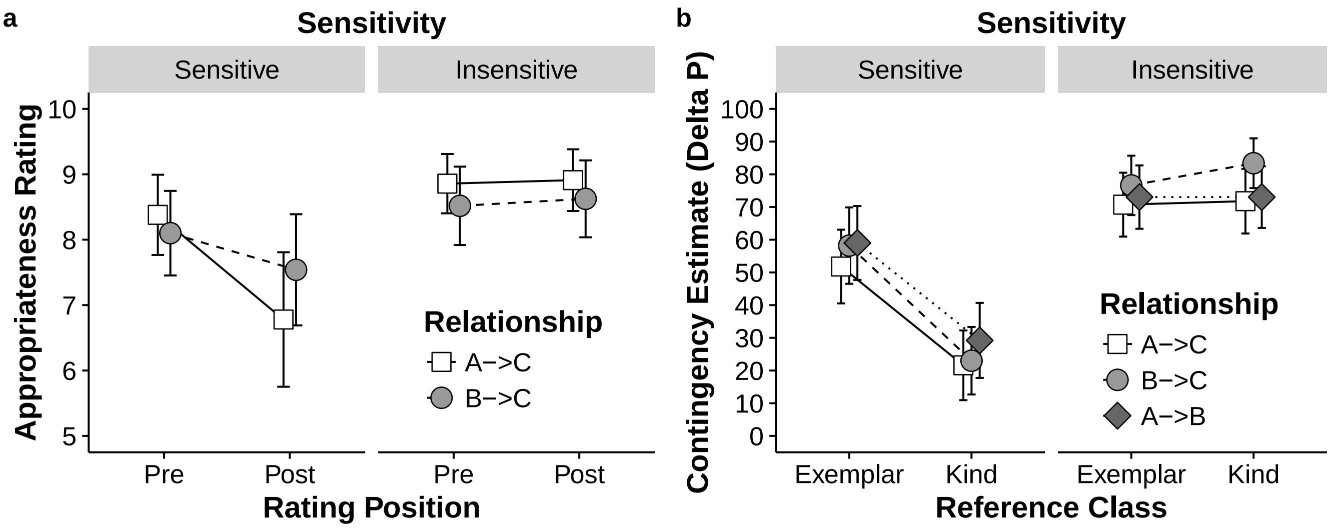

Figure 2: Group means (error bars = 95% CI) of appropriateness ratings (a) and contingency estimates (b) in Experiment 1.

marginally significant three-way Condition × Position × Relationship interaction, $F_{1,104} = 3.64$, $p < .06$, $\eta^2_g = .003$, indicating that the selective decrease of ratings from pre to post in the sensitive condition was more pronounced for the $A \rightarrow C$ relationship than for the $B \rightarrow C$ relationship, as predicted by our account.

To assess subjects' contingency estimates for the three relationships, we calculated each subject's ΔP estimate for each relationship both within Fish #1 and across the whole swarm of fish. For example, the contingency estimate for the $A \rightarrow C$ relationship within Fish #1, $\Delta P(A \rightarrow C)_{\text{Exemplar}}$, was calculated by subtracting each subject's $P(C|\neg A)_{\text{Exemplar}}$ rating from the same subject's $P(C|A)_{\text{Exemplar}}$ rating. The contingency estimates across the whole swarm were calculated analogously using the conditional probability judgments for the whole swarm. The descriptive results are summarized in Figure 2b. Most importantly, the $\Delta P(A \rightarrow C)_{\text{Exemplar}}$ estimates in the sensitive condition were not lower than the $\Delta P(B \rightarrow C)_{\text{Exemplar}}$ estimates. This finding rules out the alternative explanation that the selective drop in the $(A \rightarrow C)_{\text{post}}$ appropriateness ratings in the sensitive condition results from selectively decreased dependency estimations within Fish #1 for this particular relationship.[1]

Discussion

In this experiment, we have demonstrated that sensitive mechanisms reduce the explanatory relevance of distal causes in causal chains. Our subjects first learned that, within the narrow context of a single exemplar of a biological kind, cause A deterministically produced effect C, and that this causal relationship was always mediated by mechanism B. At this point, they interpreted the B to be a mediator of the observed $A \rightarrow C$ relationship and correspondingly found that A and B were equally crucial for the occurrence of C. However, if they subsequently found out that cause A failed to activate mechanism B in all other exemplars of the same kind, this affected their representation of the perfect dependency relation within the initially observed exemplar. Even within this exemplar, they now judged A to be less crucial for the occurrence of C than the more proximal B, indicating that A became deficient as explanation for C, despite the perfect $A \rightarrow C$ dependency relation that was observed throughout for this exemplar.

Experiment 2

In the second experiment, our main goal was to replicate the findings from Experiment 1 in the domain of artifacts. Furthermore, we made the contents of causes A and B more similar to each other in order to rule out alternative explanations for their differential treatment in the post-ratings. In Experiment 1, A was a simple, physical variable external to the system under study, while B was a complex, physiological variable internal to the system. In Experiment 2, we used only internal, physical variables that varied in complexity. We counterbalanced the assignment of the simple and the complex variable to the positions of distal cause A and proximal cause B. We expected the same pattern of results as in Experiment 1 under both assignments, showing reduced explanatory relevance of the distal cause results from mechanism sensitiv-

[1] The surprisingly low $\Delta P(B \rightarrow C)_{\text{Kind}}$ estimate in the sensitive condition resulted from very low $P(C|B)_{\text{Kind}}$ ratings. The observation that in Fish #1 (the only exemplar in which brain activity was ever recorded) brain activity reliably led to antenna illumination did not suffice to make subjects generalize this relationship across the whole kind. Hesitance to generalize from sparse data, however, is different from gathering positive evidence for sensitivity. The finding that subjects continued to regard brain activity as the crucial explanatory factor within Fish # 1 despite low kind-general contingency estimates thus does not directly disprove our hypothesis, but is certainly a finding that needs further investigation.

ity per se, regardless of its surface characteristics.

Participants

We recruited and compensated 331 subjects as in Experiment 1. 115 (35%) were removed prior to analysis according the same criteria as above. The mean age of all included subjects ($N = 216$, 109 women) was 40 years ($SD = 8.36$).

Design, Materials, and Procedure

Subjects were randomly allocated to one of four conditions that resulted from a 2 (Sensitivity: Sensitive vs. Insensitive) × 2 (Complexity of Mechanism: Complex vs. Simple) between-subjects design. They encountered an exemplar of an unknown machine, "Machine #1", with three visible devices: a single rack wheel, a complex system of rack wheels, and a fan (see Figure 3). Subjects in the complex mechanism condition saw the three devices in the arrangement shown in Figure 3, while for subjects in the simple mechanism condition, the positions of the single wheel and the complex system of wheels were reversed. The procedure was analogous to Experiment 1. In a first learning phase subjects set the leftmost device in motion and observed that this was followed by movement of the device in the middle, which was in turn followed by movement of the fan on the right. Switching off the leftmost device resulted in subsequent inertia of all variables. On the next screen, we assessed appropriateness ratings $(A \rightarrow C)_{\text{pre}}$ and $(B \rightarrow C)_{\text{pre}}$ as in Experiment 1. In the second learning phase, subjects were shown Machine #1 again together with five additional machines with identical surface features, consecutively labelled "Machine #2" to "Machine #6". They could separately intervene on each machine's leftmost device as often as they wished. In the Insensitive condition, all six machines behaved just as Machine #1 in the first learning phase. In the Sensitive condition, Machine #1 also worked just as before, but in none of the additional machines did the device in the middle or the fan ever turn on. On the following screens, subjects again indicated their $(A \rightarrow C)_{\text{post}}$ and $(B \rightarrow C)_{\text{post}}$ appropriateness ratings as well as their exemplar-specific and kind-general conditional probability judgments analogous to Experiment 1.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 4. We conducted a four-way 2 (Sensitivity: Sensitive vs. Insensitive, between-subjects) × 2 (Relationship: $A \rightarrow C$ vs. $B \rightarrow C$, within-subject) × 2 (Rating Position: Pre vs. Post, within-subject) × 2 (Complexity of Mechanism: High vs. Low, between-subjects) mixed ANOVA. We again obtained a significant two-way Sensitivity × Position interaction, $F_{1,212} = 13.02$, $p < .001$, $\eta_g^2 = .01$, indicating that the information about the additional machines affected appropriateness ratings in the sensitive condition more than in the insensitive condition. More importantly, this interaction was again qualified by a significant three-way Sensitivity × Position × Relationship interaction, $F_{1,212} = 12.23$, $p < .001$, $\eta_g^2 = .003$, indicating that the selective decrease of ratings from pre to post in the sensitive condition was more pronounced for the $A \rightarrow C$ relationship than for the $B \rightarrow C$ relationship, just as in Experiment 1. The assignment of the single wheel and the complex system of wheels to distal cause (A) vs. proximal cause (B) did not affect the results, nor did this factor interact with any of the other variables in the design. The data shown in Figure 4 is therefore collapsed across this factor.

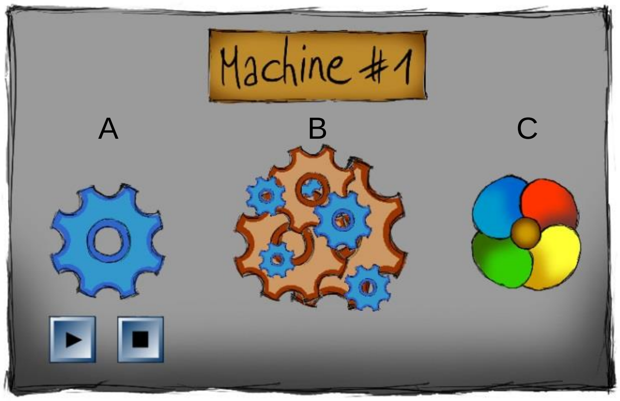

Figure 3: Screenshot of the animation used in the first learning phase of Experiment 2. Devices A and B were reversed in the Simple Mechanism condition. Letters A, B, and C were not shown to participants.

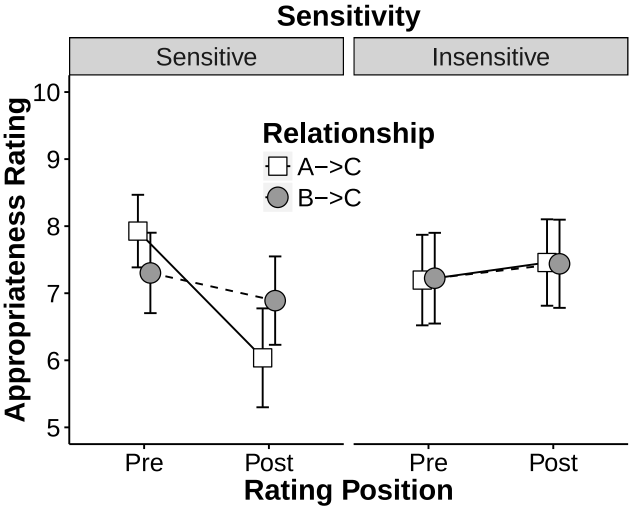

Figure 4: Group means (error bars = 95% CI) of appropriateness ratings in Experiment 2.

Subjects' contingency estimates were analyzed analogous to Experiment 1 and yielded analogous effects. The $\Delta P(A \rightarrow C)_{\text{Exemplar}}$ estimates in the sensitive condition were again as high as the $\Delta P(B \rightarrow C)_{\text{Exemplar}}$ estimates. As before, this rules out that the selective drop in $A \rightarrow C$ appropriateness ratings results from decreased $A \rightarrow C$ dependency estimates within the focal entity.

Discussion

In this experiment, we have shown that the finding that sensitive mechanisms lead to devaluation of distal causes generalizes to the domain of artifacts. If setting a physical device in motion (A) leads to movement of a second device (B), which in turn sets in motion a third device (C), both A and B are seen as equally crucial for the movement of C. However, if it is later learned that this dependency does not generalize to other exemplars of the same kind of machine, people revise their interpretation of the chain even within the first exemplar. They now judge the distal cause to be less crucial for the occurrence of the effect than before, and as less crucial than the more proximal cause that mediates the relationship. The effects were even cleaner than in Experiment 1, despite the fact that we made causes A and B more similar to each other and even counterbalanced their contents. This supports our hypothesis that high sensitivity of mechanisms *per se* leads to an interpretation of the mechanism as alternative explanation rather than as a mediator of the original relationship.

General Discussion

In two experiments, we have demonstrated that sensitive mechanisms tend to be seen as alternative explanations of their effects rather than as mediators of the causal influence of a distal cause. When people learn about a new indirect causal relationship in a single context, their default intuition seems to be that the distal cause is a crucial contributor to the terminal effect. However, if they afterwards realize that the mechanism generating the dependency between distal cause and terminal effect breaks down in most other similar contexts, they revise this intuition and devalue the causal contribution of the distal cause. The information that the mediating mechanism requires highly specific, uncommon boundary conditions makes clear that the observed A → C dependency is highly sensitive (Woodward, 2006). Sensitive causes, in turn, tend to be regarded as somewhat deficient, presumably because they fail to support future predictions and interventions in similar cases (Lombrozo, 2010; Lombrozo & Carey, 2006). As Garfinkel (1981) put it, if we want to explain the occurrence of a particular outcome, our real object of explanation is never just the occurrence of *that particular* outcome. Instead, we search for stable causes that explain the occurrence of a whole equivalence class of inessentially different outcomes. If we find out that a mechanism relates a cause to a to-be-explained effect in only a small subset of cases within the relevant equivalence class (e.g., it only works reliably in a small number of exemplars of an otherwise apparently homogeneous kind), the cause does not provide a stable explanation for this kind of effect. The discovered mechanism then turns into an alternative explanation for the outcome that screens off the influence of the distal cause from the explanandum. This explanation captures not only the present data, but also previous findings in which sensitivity of the mechanism was not directly manipulated in statistical terms, but rather implied by qualitative characteristics of the described mechanism (e.g., moral abnormality; see Nagel & Stephan, 2015).

The fact that sensitive mechanisms lead to a decrease in explanatory relevance of the distal cause A (rather than to an increase in relevance of the proximal cause B) suggests that the following psychological process might underlie the observed phenomenon. When sensitive mechanisms are observed, it becomes necessary to assume the influence of an additional, latent variable that interacts with the distal cause A to produce proximal cause B in a few but not all contexts (e.g., an abnormal preference structure of the teacher, or a genetic abnormality in Fish #1) in order to capture the structure of the complete situation. The apparent necessity of this additional variable for producing B (and, hence, C) makes it obvious that A is not sufficient in producing B (and, hence, C) even within the focal entity. Sufficiency, in turn, has been shown to be closely linked to explanatory relevance (e.g. Hilton, McClure, & Sutton, 2009; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). Future studies might aim to test this hypothesis more specifically.

Acknowledgments

We thank York Hagmayer, Jana Samland, and Michael Waldmann for helpful discussions.

References

Garfinkel, A. (1981). *Forms of explanation: Rethinking the questions in social theory*. New Haven: Yale University Press.

Hilton, D. J., McClure, J., & Sutton, R. M. (2009). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*, 383–400.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*, 942–951.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*, 167–204.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 1–25.

Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1691–1697). Austin, TX: Cognitive Science Society.

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

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*, 1–50.