Mediators or Alternative Explanations: Transitivity in Human-Mediated Causal Chains

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

Abstract

We investigate how learning that an established type-level causal relationship is implemented by human agency affects people's conceptualization of this relationship. In particular, we ask under what conditions subjects continue to perceive the original root cause as appropriate explanation for the resulting effect, and under what conditions they perceive the mediating intentional action as alternative explanation instead. Using a new experimental paradigm, we demonstrate that mechanisms involving intentional action lead to intuitions of causal intransitivity, but only when these actions are norm-violating. Potential generalizations and implications for scientific theory construction are discussed.

Keywords: explanation; causal mechanisms; causal chains; transitivity; mediators

Suppose you make the observation that in a class of pupils, all girls get high grades in a gym class, while all boys get low grades. You get the impression that gender might be an important factor to explain the grades in this class: being a girl seems to be causally relevant for a pupil to get high grades. Suppose further investigations bring to light how this causal relationship is brought about: being a girl causes you to have more flexible joints, which in turn causes you to perform better in the gym class. Given this information, does being a girl still explain why someone gets high grades? Intuitively, it does: the causal relationship of gender on grades is *implemented* or *mediated* by a physiological mechanism involving agility; the mechanism information just specifies how exactly the causal influence of gender on grades is realized.

Contrast this to your intuitions towards the following alternative information about how exactly gender influences grades in the observed class: the gym teacher likes girls better than boys and gives high grades to everyone he likes. In other words, being a girl causes you to be liked by the teacher, and being liked by the teacher causes you to receive a high grade. Intuitively, gender suddenly seems less relevant in explaining the high grades. The teacher's judgments seem to be an *alternative* explanation for the grades rather than a characterization of how exactly gender exerts its causal impact on the grades.

This example illustrates that there are different possible conceptualizations of causal mechanisms. Some intermediate causes in chains are seen as *mediators* explaining how exactly the root cause brings about its effect, while other intermediate causes are seen as *alternative explanations* for the effect in question, replacing the root cause as appropriate explanation. Furthermore, the example indicates that both intuitions can be triggered for one and the same cause-effect relationship (gender influencing grades) and with constant structural parameters (e.g., objective contingencies between the involved variables), depending only on the content of the mechanism that turns out to realize the relationship in question.

In the present paper, we will begin to investigate the psychological mechanisms that might bring about this switch in intuitions. We will first conceptualize the issue as causal transitivity problem and relate it to previous work in the field. We then turn to the question whether causal mechanisms involving intentional actions of human agents might lead to intuitions of causal intransitivity. The first experiment introduces a new paradigm designed to demonstrate the existence of the different intuitions towards the introductory example in laypeople. The second experiment tests two hypotheses as to what aspects of intentional action lead to intuitions of causal intransitivity using a different cover story. In the General Discussion, we point at implications that our results might have for transitivity intuitions outside the narrow domain of mechanisms involving intentional agents.

Transitivity in Causal Chains

Normatively, this issue can be conceptualized as the question of transitivity in causal chains. According to most classical accounts of probabilistic causality (e.g., Eells, 1991) and to the calculus of Bayesian networks (Pearl, 2000), principally, if A causes B and B causes C, then it follows that A causes C. Learning about the mechanism that implements a known causal relationship between A and C should therefore not affect the assessment of this known relationship. Technically, we can ask whether we should hold the value of B fixed when assessing the causal impact of A on C. The answer is no because B is not causally independent of A (see Rehder, 2014). For many examples, this is intuitively clear: if I want to assess whether my drinking four pints in the evening (A) causes me to have a headache on the next day (C), I should not hold fixed the amount of toxins produced in my body in the meantime (B). Causality is transitive: A *per se* is seen to be critical for C even when it is established that its causal influence is entirely mediated via B—as in the agility version of the introductory example.

However, the intuition that causality is transitive in causal chains is not always observed in the causal judgments of laypeople. Recently, Johnson and Ahn (2015) presented subjects with descriptions of numerous token causal chains (e.g., "Allison exercised for 20 min [A], then Allison became thirsty [B], then Allison drank a whole bottle of Water [C]")

and then asked them to what extent they would say that A caused B, to what extent B caused C, and to what extent A caused C. For some examples, like the one above, they found high ratings for all three causal relationships, indicating transitivity. For other token chains, however, causality was not seen as transitive. For example, in "Ned ate very spicy food [A], then Ned drank a lot of water [B], then Ned had to urinate [C]", people rendered high causality ratings for $A \rightarrow B$ and for $B \rightarrow C$, but much lower ratings for $A \rightarrow C$. Thus, even though both links of the chain were seen as highly causal, the root cause was judged to have only a low causal impact on the final effect. This is analogous to our intuitions towards the teacher version of the introductory example: despite the fact that the teacher's sympathy (B) is causally dependent on the pupil's gender (A), it seems that B is seen as an alternative cause of the grades (C) rather than as a descendant of A implementing the mechanism leading from A to C.

Note that in these cases it is not denied that A causes B. The data by Johnson and Ahn (2015) indicate that people can be fully aware of this relationship, yet see B in some sense as an independent cause of C (as it can be causal for C where A is not). Thus, the judgment that A *per se* is not causal for C does not reflect a belief that there is no (indirect) causal connection between A and C. Rather, it seems to reflect the intuition that A's causal impact on C is mediated via the "wrong" kind of mechanism. In other words, there seem to be some mechanisms which are compatible with the notion that the root cause *per se* matters for the effect, while there are other mechanisms which are incompatible with this notion.

Johnson and Ahn (2015) used token causal chains describing everyday actions of individuals and their effects. Causal transitivity varied widely across their individual items (see examples above). They showed that the extent to which such chains are causally transitive is a function of the extent to which the chains are represented in semantic memory as one single chunk (rather than as two separate relationships that are stored independently). However, the data do not allow conclusions as to which *item* properties cause a chain to be represented one way rather than the other. It thus seems unclear how their account could handle our intuitions towards the introductory example, where new mechanism knowledge is discovered for a constant unfamiliar type-level causal relationship which is unlikely to have a pre-established representation in semantic memory. Additional processes seem to be at work here.

Intentional Agents Implementing Causal Chains

Which features of a discovered mechanism determine the resulting transitivity intuitions? A salient property of the teacher mechanism in the introductory example is that it involves an intentional agent as realizer of the causal relationship in question, while the agility mechanism is part of a blind biological process. Previous research suggests that intentional actions are particularly likely to be selected as the principal cause of terminal effects in unfolding token causal chains. Hilton, McClure, and Sutton (2009) have shown that when asked to identify the actual cause of a token event, people trace back the antecedent causal chain until they reach an intentional agent and designate his action to be the principal cause of the effect—even if the downstream causal process involves highly abnormal events that would be designated to be the principal cause in the absence of upstream intentional agency (see also Hilton & Slugoski, 1986). In other words, intentional root causes seem to make chains transitive even when they involve highly abnormal events. Accordingly, one may suspect that the reverse might also hold: finding out that a physical root cause influences its effects by affecting intentional agents' decisions may make the chain *in*transitive. Intentional agents may be generally seen as unmoved movers that initiate causal processes rather than merely transfer external influences, producing intransitive chains and screening off the influence of the root cause from the explanandum.

However, in the materials used by Hilton et al. (2009), the intentional actions that were selected as explanations for their distal effects were usually also morally wrong or at least highly negligent. The same is true for the teacher's grading practice in the introductory example which obviously involves morally dubious criteria. According to Hitchcock and Knobe (2009), morally abnormal actions tend to be selected as causes in common-effect structures. Rather than for their status as intentional actions, the explanations in the causal chains in Hilton et al. (2009) may have been selected for their status as especially abnormal events. In this case, intentional actions that are not norm-violating should not be seen as alternative explanations relative to the antecedent cues by which they are triggered, but as proper mediators instead. Experiment 2 is designed to differentiate between the intentionality and the abnormality hypotheses.

For our experiments, we did not employ a causal selection task for the explanation of a single mundane token event. Rather, we wanted to explore whether the conceptualization of one and the same type-level causal relationship can be differentially affected by learning that it is implemented by different kinds of causal mechanisms. In our experimental paradigm, we first teach all subjects the existence of a type-level causal relationship ($A \rightarrow C$) and assess (i) how appropriate it would be to state that A is crucial for C. We then provide different groups of subjects with different information about the mechanism implementing this relationship ($A \rightarrow B \rightarrow C$, where the content of B is varied between subjects). Afterwards we assess once again how appropriate it would now be to state (ii) that A is crucial for C and (iii) that B is crucial for C. If the second rating for A is as high as the first rating for A, this will indicate that the mechanism elicited a transitivity intuition: the fact that A causes C *via B* is compatible with the notion that A *per se* matters for C. By contrast, if the rating for A is decreased in response to learning about a specific mechanism while the rating for B is at least as high as the first rating for A, this will indicate that the chain is seen

to be intransitive: the intermediate cause is conceptualized as an *alternative explanation*, replacing A as appropriate explanation for C.

Experiment 1

In the first experiment we sought to establish that the paradigm outlined above is able to capture the intuitive differences displayed in the introductory example.

Participants

The experiment was conducted as an online study. A total of 171 subjects from the UK completed the experiment, 32 of which were excluded from the statistical analyses because they failed in a simple attention check question that we asked at the end of the experiment. The average age of all included subjects (N = 139, 93 women) was 38 years (SD = 8.62).

Design, Materials, and Procedure

We constructed a complete 2 (Mechanism: Physiology vs. Teacher) × 2 (Contingency: Deterministic vs. Probabilistic) × 2 (Balance: Boys with high grades vs. Girls with high grades) between-subjects design. Subjects in all conditions were asked to take the perspective of a scientist investigating the relationship between pupils' gender (A) and their grades in a physical education class (C). In a first learning phase they received data of a class of ten pupils (five boys and five girls) in tabular form which showed each pupil's gender (A vs. ¬A) and whether he or she got a high grade (C) or a low grade (¬C). Whether being a boy or a girl was designated as A was counterbalanced with the Balance factor. When boys had high grades, the grades in question were for a course in athletics; when girls had high grades, they were for a course in gymnastics. Half of the subjects learned that there was a deterministic relationship between gender and grades, for the other half this relationship was probabilistic (one exception for each gender). This contingency manipulation was included to explore whether intransitivity intuitions can be elicited for both deterministic and probabilistic causal relationships. After having studied this table, participants were asked to indicate on an 11-point scale (ranging from 0 to 10) how appropriate they found the following sentence to describe the concrete observations they have just made: "The gender of a pupil is crucial for this pupil's grade in athletics/gymnastics" (we call this *appropriateness rating* $[A \rightarrow C]_{pre}$ because it is measured prior to the introduction of mechanism information). At this point, we expected that subjects would have formed the impression that, within the observed sample, gender influences grades (A → C), leading them to render unanimously high appropriateness ratings.

The crucial mechanism manipulation was introduced in a second learning phase. Subjects were told that they would have come up with a hypothesis about the underlying mechanism. Half of them (Mechanism: Physiology) were told that they suspected boys to develop higher muscularity than girls which in turn would lead them to receive higher grades in athletics. (In the other Balance condition, girls were suspected to develop higher agility than boys which in turn would lead them to receive higher grades in gymnastics.) The other half (Mechanism: Teacher) was told that they suspected the teacher to like boys better than girls (and vice versa for the other Balance condition), leading boys (girls) to receive higher grades. In all conditions, subjects read that they went back to collect additional data from the same class corresponding to their hypothesis. These data were then presented in a new version of the table which now included an additional column representing each pupil's value on the suspected mediating variable (B [high] vs. ¬B [low]; in both Contingency conditions, this new variable deterministically predicted the grades). After having studied this extended table, subjects were again asked to indicate how appropriate it would be to state that A is crucial for C ($[A \rightarrow C]_{post}$), and also how appropriate it would be to state that B is crucial for C ($[B \rightarrow C]_{post}$) using the same question format (B being described as "the muscularity of a pupil/the agility of a pupil/the teacher's sympathy for a pupil", depending on condition). We expected continuously high ratings for $(A \rightarrow C)_{post}$ in the Physiology condition and a drop in ratings for $(A \rightarrow C)_{post}$ in the Teacher condition.

Afterwards, we assessed contingency estimates for all three relationships from memory (i.e., P[C|A] vs. P[C|¬A], P[B|A] vs. P[B|¬A], and P[C|B] vs. P[C|¬B]), assessed on six separate 11-point scales ranging from 0 (impossible) to 100 (certain) to see if subjects in both Mechanism conditions based their judgments on the same impression of the objective probabilities. In the Probabilistic conditions, we furthermore asked them to indicate the conditional dependence of C on A given constant values of B (i.e., P[C|A∧B] vs. P[C|¬A∧B], and P[C|A∧¬B] vs. P[C|¬A∧¬B]) to see if they understood that, in the observed sample, C was independent of A when B was held fixed (regardless of the Mechanism condition). Finally, we wanted to know how plausible they found each of the described causal relationships (A → C, A → B, and B → C) according to their prior real world knowledge, regardless of the fictional data from the experiment.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 1. We conducted a global 2 (Mechanism: Physiology vs. Teacher) × 2 (Contingency: Deterministic vs. Probabilistic) × 2 (Balance: Boys high vs. Girls high) × 3 (Rating: $(A \rightarrow C)_{pre}$ vs. $(A \rightarrow C)_{post}$ vs. $(B \rightarrow C)_{post}$, within-subject) mixed ANOVA. Since there was neither a main effect of Balance, $F_{1,131} < 1$, nor any significant interaction effect including Balance, largest $F_{2,262} = 2.19$, data in Figure 1 are averaged across this factor. There was a main effect of Rating, $F_{2,262} = 22.00$, $p < .001$, $\eta_p^2 = .14$, and a significant interaction of Rating × Contingency, $F_{2,262} = 5.99$, $p = .003$, $\eta_p^2 = .04$. There was also a trend for an interaction of Rating × Mechanism, $F_{2,262} = 2.87$, $p = .06$, $\eta_p^2 = .02$.

The main prediction that we made was that we would see a distinct drop of appropriateness ratings for $(A \rightarrow C)_{post}$ relative to $(A \rightarrow C)_{pre}$ selectively within the Teacher conditions.

A planned contrast testing whether this difference was larger in the Teacher compared to the Physiology condition confirmed this prediction, $t_{131} = 2.97$, $p < .05$, $r = .25$. Furthermore, the ratings for $(B \rightarrow C)_{post}$ were at least as high as the ratings for $(A \rightarrow C)_{pre}$ in all conditions (see Figure 1). The increase in appropriateness ratings for $(B \rightarrow C)_{post}$ in the Probabilistic conditions can be explained by the fact that B is a deterministic predictor for C while A is not (i.e., the probabilistic element lies in $A \rightarrow B$).

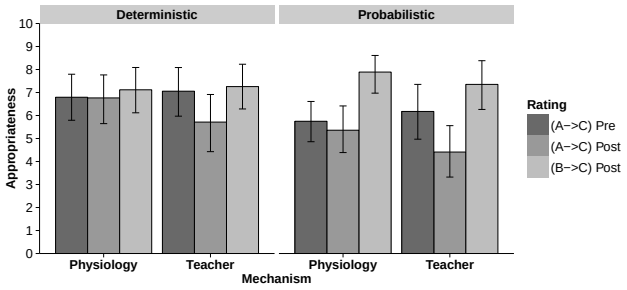

Figure 1: Group means (Errorbars = 95% CI) of appropriateness ratings in Experiment 1.

An analysis of the subjects' contingency estimates showed that, globally, subjects represented the contingencies contained in the learning data quite adequately, apart from generally underestimating the strength of the deterministic relationships. Crucially, this pattern was similar in both the Physiology and the Teacher condition. Furthermore, participants in the Probabilistic conditions recognized that A and C were conditionally independent given constant levels of B ($P[C|A \wedge B] \approx P[C|\neg A \wedge B]$ and $P[C|A \wedge \neg B] \approx P[C|\neg A \wedge \neg B]$). This pattern was also consistent in both Mechanism conditions. The pattern of contingency estimates therefore cannot account for the differences we observed for the appropriateness ratings between the Mechanism conditions. The same holds for the plausibility ratings: the relationships $A \rightarrow B$ and $B \rightarrow C$ were rated to be equally plausible for both mechanisms.

Discussion

The results of this experiment show that acquiring different mechanism knowledge can differentially affect the conceptualization of a given type-level causal relationship. When effects of gender on grades in a gym class were mediated via a physiological process, this was seen as compatible with the notion that gender *per se* matters for the grades. This was not the case when the relationship was mediated via the teacher's personal preference for the pupil. In this case, subjects revised their earlier judgment that gender was critical for the grade and attributed the grades to the teacher's sympathy instead, even though they were aware that the teacher's preference was causally affected by the pupils' gender. This result cannot be explained by differences in encoded contingencies or prior plausibility intuitions for the different mechanisms.

Experiment 2

In Experiment 2, we attempted to replicate the phenomenon using a more artificial cover story. This story allowed us to differentiate between the intentionality and the abnormality hypotheses developed in the theory section: is the root cause always screened off when the mechanism involves an intentional agent, or only when this agent's intentions are morally dubious?

Participants

The experiment was conducted as an online study. We recruited 232 subjects from the UK, 31 of which failed in the attention test and were excluded from all analyses. The average age of all included subjects ($N = 201$, 118 women) was 36 years ($SD = 8.14$).

Design, Materials, and Procedure

We used the same experimental paradigm as in Experiment 1. This time, we implemented four between-subjects conditions describing different mechanisms underlying the same causal relationship. Subjects in all conditions were asked to imagine they were at a funfair where they were observing a swing tossing passengers around. They read that sometimes after a passenger had entered the swing a red flashlight came on, whereupon the passenger had to leave the swing without having taken a ride. Participants read that they would have gained the impression that the flashlight (C) came on more frequently for corpulent passengers (A). Upon studying the learning data about a deterministic relationship between A and C across ten observed passengers (the first five of them corpulent, the last five not corpulent), they rendered their $(A \rightarrow C)_{pre}$ appropriateness rating.

We then told subjects in the different conditions that they would have come up with one of four hypotheses about the underlying mechanism. In the first condition (Scale) they were told that they suspected the flashlight to be part of a safety mechanism. They believed a scale to be built into the swing which causes the flashlight to come on whenever the loading exceeds a critical threshold. This condition was intended to provide a baseline for unequivocal judgments of transitivity, analogous to the Physiology condition in Experiment 1. In the second condition (Accurate Operator) the subjects' hypothesis was that the operator intends to guarantee the safety of his costumers, and that whenever he judges a passenger to be too corpulent for his swing he activates the flashlight. This condition was intended to provide a mechanism involving an intentional agent but otherwise being closely matched to the Scale condition. According to the intentionality hypothesis, this chain should nonetheless be seen as intransitive, while it should be seen as transitive according to the abnormality hypothesis since the operator does not violate a norm. The mechanism hypothesis of the third condition (Biased Operator) was that the operator does not like corpulent people and enjoys embarrassing them, and that whenever he judges a passenger to be too corpulent for his taste, he activates the flashlight. This condition was intended as providing

an intentional mechanism analogous to the Teacher condition in Experiment 1 which should elicit judgments of intransitivity according to both the intentionality and the abnormality hypothesis. In the last condition (Central Computer), the subjects' hypothesis was that a central computer supervising the electricity supply detects irregularities at unpredictable times, and then generates an electronic signal that causes the flashlight to come on. The co-occurrence of corpulence and flashlight in the observed sample would have resulted from mere coincidence. This condition was intended to create a structure in which A turned out to be actually causally irrelevant for C. B (the central computer) caused the effect, and the relationship between A and C in the sample was merely due to a coincidental correlation between A and B in the observed sample. This condition thus provides a baseline for an unequivocal alternative explanation. The crucial question is where intuitions towards the mechanisms involving intentional operators fall within the space that is spanned between Scale (a clear mediator) and Central Computer (a clear alternative explanation).

In all conditions, subjects read that they went over to the operator and asked him for information concerning the mechanism. The operator confirmed their hypothesis in all conditions and told them for the last ten passengers whether or not the scale had detected a threat to the passenger/he had detected a threat to the passenger/he had not liked the passenger/the central computer had detected irregularities in the electricity supply (depending on condition). After having studied the corresponding extended table in which all three relationships were deterministic, subjects again rendered their appropriateness ratings for $(A \rightarrow C)_{post}$ and $(B \rightarrow C)_{post}$. Finally, subjects indicated their contingency estimates for all three relationships from memory. Plausibility ratings were not collected.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 2. We conducted a global 4 (Condition: Scale vs. Accurate Operator vs. Biased Operator vs. Central Computer) $\times$ 3 (Relationship: $(A \rightarrow C)_{pre}$ vs. $(A \rightarrow C)_{post}$ vs. $(B \rightarrow C)_{post}$, within-subject) mixed ANOVA. There was no main effect of Condition, $F_{3,\,197} = 1.81$, but a main effect of Relationship, $F_{2,\,394} = 7.48$, $p < .001$, $\eta_p^2 = .04$. The global Relationship $\times$ Condition interaction was not significant, $F_{6,\,394} = 1.74$, $p = .11$, $\eta_p^2 = .03$. However, planned contrasts revealed that the decrease in appropriateness ratings from $(A \rightarrow C)_{pre}$ to $(A \rightarrow C)_{post}$ did not differ between Scale and Accurate Operator, $t_{197} < 1$, but was larger compared to Scale both in the Biased Operator condition, $t_{197} = 2.41$, $p < .05$, $r = .17$, and in the Central Computer condition, $t_{197} = 2.81$, $p < .01$, $r = .20$. Furthermore, the decrease was as large in Biased Operator as in Central Computer, $t_{197} < 1$. At the same time, $(B \rightarrow C)_{post}$ was not smaller than $(A \rightarrow C)_{pre}$ in any of the conditions, largest $t_{197} = 1.48$.

The contingency estimates were similar across all four mechanism conditions for all three causal relationships

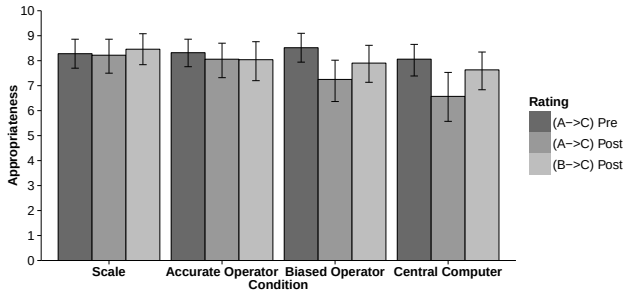

Figure 2: Group means (Errorbars = 95% CI) of appropriateness ratings in Experiment 2.

($A \rightarrow B$, $B \rightarrow C$, $A \rightarrow C$). The only exception was that ΔP for $A \rightarrow B$ tended to be higher in the Scale condition than in the other three conditions, a difference that was significant in the comparison with Biased Operator, $t_{197} = 2.28$, $p < .05$. This indicates that the contingency between corpulence and the biased operator's judgment was perceived to be weaker than the contingency between corpulence and threat detection by the scale, despite identical learning data. While this difference may add to the decrease in $(A \rightarrow C)_{post}$ in the Biased Operator condition, the overall pattern of contingency estimates cannot fully account for the overall pattern of appropriateness ratings.

Discussion

The results of Experiment 2 demonstrate that involving an intentional agent is not a sufficient property for a mechanism to elicit judgments of intransitivity. Agents intending to accurately transfer an objective signal from A to C seem to be conceptualized akin to a mechanical process serving the same function. However, if the same relationship is dependent on an agent's highly idiosyncratic (and morally dubious) preference structure, the physical root cause is screened off from the explanandum to the same extent as if it was merely a coincidental, causally irrelevant confound.

General Discussion

In two experiments, we have demonstrated that the conceptualization of one and the same established type-level causal relationship can be differentially affected when knowledge about different causal mechanisms is acquired. Some mechanisms (e.g., physiological processes) are compatible with the notion that the root cause *per se* matters for the effect, while others (e.g., biased judges) provide an *alternative* explanation, leading the root cause to be seen as a less appropriate explanation for the effect. Our data indicate that these differences are not brought about by different causal strength assessments, nor by different plausibility intuitions. Also, involving an intentional agent does not constitute a sufficient condition for a mechanism to elicit intuitions of intransitivity. In the contexts we investigated, this intuition seems to depend crucially on the intentional agent being morally abnormal.

This latter finding may indicate that a more general psychological mechanism may underlie our findings which might make our framework applicable to intransitivity intuitions outside the narrow scope of mechanism involving human agents. Immoral human agents may merely be a very salient instance of abnormal mechanisms more generally. Recently, Kominsky, Phillips, Gerstenberg, Lagnado, and Knobe (2015) have argued that moral and statistical abnormality of potential causes function similarly in eliciting counterfactual possibilities that in turn have similar downstream effects of the assessment of other causes in the same common-effect network. It is possible that similar generalizations hold in our case. In the Physiology condition from Experiment 1, the mechanism is implemented in every single pupil in a lawlike manner. Thus, it can be assumed that an $A \rightarrow C$ relationship brought about by this mechanism will be stable across most perturbations of the entity implementing the mechanism in each token case, both within and beyond the observed sample. Contrast this with the teacher example: the observed type-level $A \rightarrow C$ relationship is entirely dependent on this particular teacher implementing the mechanism in the observed sample. Replacing the teacher with pretty much any colleague would presumably make the $A \rightarrow C$ relationship disappear. Even though A is doubtlessly an indirect cause of C in the observed sample, explaining C in terms of A might feel inadequate because the relationship can be expected to be highly sensitive to minor perturbations of the boundary conditions provided by the particular teacher implementing the mechanism in the observed sample (see also Garfinkel, 1981).

So far, these are only speculations as to how far our findings may generalize which still need to be empirically tested with materials outside the domain of human agency. In case of success, this account may be applicable not only to aspects of everyday causal cognition, but even to psychological processes underlying scientific theory construction. Whenever a scientific theory is about a causal chain structure (e.g., process X influences process Y, which in turn leads to effect Z), the issue discussed in this paper arises: if the researcher wants to assess whether process X *per se* explains outcome Z, should she hold process Y constant? If she conceives of process Y as a mediator, the answer is definitely no. But there might be cases in which, despite the underlying chain structure, she conceives of process Y as an *alternative explanation* of effect Z. Sometimes, it does not seem clear which of these conceptualizations is more adequate—yet, the decision for one of them will crucially shape the methodology and conclusions of the subsequent research (e.g., decisions about whether process Y is to be controlled or to be left free to covary with X).

Acknowledgments

We thank Jana Samland, Jonathan Kominsky, Joshua Knobe, and Michael Waldmann for helpful comments.

References

Eells, E. (1991). *Probabilistic causality* (Vol. 1). Cambridge: Cambridge University Press.

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

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

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*, 75.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*, 587–612.

Johnson, S. G., & Ahn, W.-K. (2015). Causal networks or causal islands? the representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*.

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

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

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107.