Establishing backward causation on empirical grounds: An interventionist approach*

Alexander Gebharter · Dennis Graemer · Frenzis H. Scheffels

Abstract: We propose an analysis of backward causation in terms of interventionism that can avoid several problems typically associated with backward causation. Its main advantage over other accounts is that it allows for reducing the problematic task of supporting backward causal claims to the unproblematic task of finding evidence for several ordinary forward directed causal hypotheses.

Keywords: backward causation, interventionism, empirical evidence

1 Introduction

According to Faye (2018), the concept of backward causation "raises two sets of questions: those concerning conceptual problems and those that relate to empirical or physical matters" (ibid., sec. 2). Conceptual problems include bootstrap

*This is a draft paper. The final version of this paper is forthcoming under the following bibliographical data: Gebharter, A., Graemer, D., & Scheffels, F. H. (forthcoming). Establishing backward causation on empirical grounds: An interventionist approach. *Thought: A Journal of Philosophy*.

paradoxes (cf. Mellor, 1991), consistency paradoxes (cf. Lewis, 1976), and the bilking argument (cf. Black, 1956). Empirical and physical matters comprise the problem of finding a physical realizer for backward causal relations and the question of how to establish backward causal hypotheses. In this paper we cover both kinds of questions to some extent. After introducing Woodward's (2003) interventionist theory of causation in section 2, we propose an interventionist analysis of backward causation in section 3. We also argue that this analysis does not run into conceptual problems such as the ones mentioned above. Then, in section 4, we turn to empirical matters: Since interventionism has no implications for how causation might be physically realized, we bracket the problem of finding a physical realizer for backward causation and rather focus on showing how the account can be used for supporting backward causal claims on empirical grounds by testing several quite harmless forward directed causal hypotheses. While interventionism is not the only theory of causation that can avoid conceptual problems, the fact that it provides the resources to empirically establish backward causal relations is what clearly sets it apart from other accounts. To further emphasize this main advantage, we compare our account to three other types of theories of causation. We conclude in section 6.

2 Interventionism

In this section we introduce the basics of interventionism required for our analysis of backward causation. Interventionism is based on the simple idea that effects can be influenced by manipulating their causes. Woodward (2003, p. 59) provides the following manipulationist characterization of causation:

> **(M)** X is a direct cause of Y w.r.t. $\mathbf{V}$ if and only if there is a possible intervention on X w.r.t. Y that will change the probability distribution of Y when all other variables $Z_i \in \mathbf{V}$ are fixed by in-

tervention. X is a contributing cause of Y w.r.t. $\mathbf{V}$ if and only if there is a directed path from X to Y such that there exists a possible intervention on X that will changes Y's probability distribution when all other variables in $\mathbf{V}$ that are not on this path are fixed by interventions.

In order to intervene on X w.r.t. Y, a so-called intervention variable is required (Woodward, 2003, p. 98):

> **(IV)** I_X is an intervention variable for X w.r.t. Y if and only if the following four conditions are satisfied:
>
> ① I_X causes X.
>
> ② Some of I_X's values (the intervention variable's *on*-values) screen X off from all its other causes Z_i.
>
> ③ If I_X causes Y, then only through X.
>
> ④ I_X is statistically independent of any cause Z_j of Y that causes Y over a path that does not go through X.

An intervention on X w.r.t. Y can then be defined as an intervention variable I_X for X w.r.t. Y taking one of its *on*-values that is associated with a change in X. The notion of an intervention variable is designed in such a way that it picks out exactly those potential causes I_X of X that can be used for testing whether X is a (direct or contributing) cause of Y w.r.t. a set of variables $\mathbf{V}$. (For details, see Woodward, 2003, sec. 3.1.4.)

Before presenting our analysis of backward causation, a few remarks on interventionism seem appropriate. Firstly, intervention variables do not have to describe human actions. Secondly, interventionism is a type-level theory introducing causation as a relation between variables. Thirdly, direct and contributing causation are characterized relative to a set of variables, but the notion of

an intervention is not relativized in that way. When speaking of causation in **(IV)**, Woodward (2003) consequently refers to what he later called causation simpliciter: X is a cause of Y simpliciter "as long as it is true that there exists a variable set $\mathbf{V}$ such that X is correctly represented as a contributing cause of Y with respect to $\mathbf{V}$" (Woodward, 2008, p. 209). Fourthly, not all four conditions in **(IV)** are actually required to account for direct and contributing causation in terms of **(M)**. As Baumgartner and Drouet (2013, pp. 186f) show, ② is actually dispensable. Finally, neither **(M)** nor **(IV)** make any reference to time, which renders the theory in principle compatible with forward, simultaneous, and backward causation. We agree with Hausman (1998) that this is a virtue of a theory of causation; in the case of interventionism it allows one to use the theory as a tool for investigating questions concerning backward causation on empirical grounds.

3 Analyzing backward causation

Let us now come to explicating backward causation within interventionism. We propose the following analysis:

> **(BC)** X is a backward cause of Y iff (i) X is a cause of Y simpliciter and (ii) Y-instantiations brought about by interventions on X w.r.t. Y occur strictly before the corresponding X-instantiations.

Condition (i) specifies the genus and condition (ii) the differentia for backward causation to the background of interventionism. **(M)** and **(IV)** guarantee that the only explanation for a change in Y induced by an intervention on X w.r.t. Y requires X to be causally relevant for Y, and (ii) that the direction of X's causal influence on Y is actually backward in time.

Note that the notion of backward causation proposed is metaphysically cau-

tious. As we will see below, it does not come with strong metaphysical commitments, which is mainly due to the fact that interventionism itself is a metaphysical lightweight account of causation. Condition (ii) is, for example, weak enough to allow for changes in Y that occur simultaneously or later than changes in X if these Y-changes are not induced by an intervention on X w.r.t. Y. As a consequence, interventionism allows—at least in principle—for temporal loops: Y might turn out to be a forward cause of X while X is a backward cause of Y. But would this not imply that the account of backward causation proposed runs directly into bootstrap paradoxes? The problem here seems to be that the cause presupposes its effect which, in turn, presupposes its cause. Firstly, note that interventionism renders the question of whether such temporal loops exist an empirical one—its answer fully depends on whether the right interventions exist. It might well be the case that they do not. Secondly, casting intuitive doubt does not suffice to constitute a paradox. For a serious problem it would be required that the possible existence of causal cycles somehow leads to inconsistencies. But this seems not to be the case in an interventionist setting. Everything required for a causal cycle between two variables X and Y is the existence of an intervention on X w.r.t. Y being associated with changes in Y's probability distribution and the existence of an intervention on Y w.r.t. X inducing changes in X's probability distribution. Such interventions clearly exist if X and Y describe, for example, the positions of two magnets on a table. This is, of course, a case of simultaneous causation. However, it shows that the consistency of this scenario to the background of interventionism does in no way depends on the times at which X- and/or Y-values are instantiated.

Note that the account proposed can also avoid consistency paradoxes. Such paradoxes typically arise when changing the past trough backward causation would result in inconsistencies. A classical example would be a case where a

person travels backward in time and kills her younger self. The problem consists in the fact that if backward causation is possible, the following two propositions seem intuitively plausible, while taken together they are inconsistent:

1. It is possible for someone to kill her younger self.

2. It is impossible for someone to kill her younger self.

Proposition 1. is plausible because nothing seems to exclude the possibility to kill one's younger self when going backward in time is possible. Killing someone is (moral issues aside) an ordinary human action like any other. And proposition 2. is plausible because killing one's younger self seems to prevent one from going back in time and killing one's younger self. Again, we favour empirically informative approaches to metaphysical issues and, thus, think that intuitive plausibility should not be considered as a reliable test for the seriousness of a problem. The implementation of backward causation into interventionism provides a much handier evaluation tool: It is easy to see that propositions 1. and 2. are not supported by an interventionist treatment of backward causation. As already said, whether backward causal relations exist (an assumption on which both 1. and 2. build) becomes an empirical question within an interventionist framework. In addition, interventionism does neither provide the resources to infer the possibility nor the impossibility for someone to kill her younger self if backward causal relations exist. The theory does not even imply that present or future events can actually change anything that already happened in the past. In other words, it does not imply that the value a variable Y has actually taken can be changed afterwards by intervening on Y's backward cause X. Strictly speaking, everything interventionism says about backward causation is that Y-instantiations induced by interventions on X w.r.t. Y occur strictly before the corresponding X-instantiations.

Finally, also the threat posed by the bilking argument can be avoided. In

a nutshell, the bilking argument says that if X is a backward cause of Y, then Y-instantiations brought about backward in time by X-instantiations can, even after they occurred, still be prevented by intervening on X. The problem is that the prevented X-instantiations cannot have caused these Y-instantiations, which contradicts the assumption that they actually did. Here comes the interventionist response: Firstly, note that the argument requires reference to variable instantiations being causally relevant for variable instantiations. It smuggles token-level causal claims into interventionism that lack a clear meaning to the background of that framework. What we can say from the viewpoint of interventionism anyway is the following: Whether X can actually be controlled by an intervention after Y has taken one of those values that depend on X-values induced by interventions on X w.r.t. Y is irrelevant for whether X is a backward (type-level) cause of Y. The only thing required for X to be a backward (type-level) cause of Y is that there exists a suitable intervention variable I_X for X w.r.t. Y such that $I = on$ induces changes in Y's probability distribution when all off-path variables are fixed by interventions and that the relevant Y-values occur strictly before the changes in X induced by $I = on$. Also note that an effect variable such as Y taking a certain value can typically have different causes. So it seems plausible to infer from the observation that Y takes one of those values it typically takes under an intervention on X w.r.t. Y and that X has been prevented to take one of those values typically associated with these Y-values by another intervention, that another cause of Y must be responsible for Y's taking this particular value (or that Y has taken this particular value by chance, if we allow for indeterministic causation). The situation is structurally identical to a scenario in which an ordinary effect Y of X takes a value y that is typically associated with values $x_1, ..., x_n$ of X brought about by an intervention on X w.r.t. Y and in which X has actually been forced to

take a value x_i (with $i > n$) by another intervention. The obvious consequence from such an observation is that Y has taken value y because of another cause (or, again, by chance).

Note that interventionism is not the only theory of causation that can avoid the conceptual problems discussed above. However, to the best of our knowledge no one has yet argued that interventionism can avoid these problems and the work needs to be done. The fact that the theory can avoid these problems is also essential for its usefulness when it comes to establishing backward causal hypotheses on empirical grounds. Would it run into conceptual problems, then the theory would not be much worth, even if it can support backward causation on the basis of empirical findings.

4 Establishing backward causation

Let us now come to what distinguishes our analysis of backward causation within interventionism from other accounts: to the question of how backward causal hypotheses can be supported on empirical grounds. Let us briefly illustrate why this task seems especially problematic. Assume we want to test the hypothesis

H: X is causally relevant for Y.

In order to test H, we bring about X-changes via experimental manipulation and check whether Y-changes occur. If we observe such Y-changes, then there are two possible explanations:

(a) H is true.

(b) Something went wrong; the experimental setup was flawed, hidden confounders were involved, etc.

Whether we consider (a) or (b) as more likely and whether we tend to take our result as evidence for H crucially depends on whether the observed Y-changes

occurred after or before the X-changes. If the former was the case and we do not have specific reasons for (b), then we tend—in accordance with scientific practice—to interpret the experiment's result as evidence for H. If the latter was the case, however, most scientists would probably shy away from such an interpretation. They would consider (b) much more likely in that case. Possible reasons for this are that backward causation is somewhat unfamiliar and has an otherworldly touch, has not been scientifically confirmed so far, is surrounded by conceptual and other problems, etc.

In this section we argue that an interventionist treatment of backward causation can help in overcoming this problem. In particular, we argue that the problematic task of testing for backward causation can be reduced to the unproblematic task of finding evidence for several ordinary forward causal hypotheses. We tried to keep the formal details as minimal as possible. However, since our argumentation crucially depends on the technical details of interventionism (as introduced in section 2), we could not avoid to go into technical details to some extent. Our result directly follows from interventionism, **(BC)**, and the classical understanding of confirmation as probability increase. Let us begin by recalling that, according to **(BC)**, X is a backward cause of Y if (i) X is a cause simpliciter of Y and (ii) Y-instantiations associated with interventions on X w.r.t. Y occur strictly before the corresponding X-instantiations. (i) is true if X is a direct or contributing cause of Y w.r.t. some set of variables $\mathbf{V}$, and according to **(M)**, this is the case if there exists an intervention on X w.r.t. Y that induces changes in Y if the values of all off-path variables in $\mathbf{V}$ are fixed by interventions. From a logical point of view, **(M)** thus explicates that X is a cause of Y w.r.t. variable set $\mathbf{V}$ via an existentially quantified conjunction of the form

$$p \leftrightarrow \exists I_X (q_{[I_X]} \wedge r_{[I_X]}), \tag{1}$$

where the different parts of Equation 1 are interpreted as follows:

p ... X is a cause of Y w.r.t. $\mathbf{V}$.

$\exists I_X(\dots)$ ... there exists an I_X such that $(\dots)$.

$q_{[I_X]}$ ... I_X is an intervention variable for X w.r.t. Y.

$r_{[I_X]}$ ... Changes in I_X are associated with changes in Y's probability distribution when all off-path variables are held fixed.

It hence follows from interventionism that once an I_X is found that satisfies the conditions ①, ③, and ④ specified in **(IV)**[1] for X w.r.t. Y ($q_{[I_X]}$) and changes in I_X are actually associated with changes in Y's probability distribution ($r_{[I_X]}$), then X is a cause of Y w.r.t. $\mathbf{V} = \{X, Y\}$ (p). The $r_{[I_X]}$-part of Equation 1 is easy to test for any candidate intervention variable I_X for X w.r.t. Y. One just has to check whether I_X and Y are correlated. The $q_{[I_X]}$-part, on the other hand, is a little bit trickier. However, establishing the $q_{[I_X]}$-part is essential because it would guarantee that nothing was wrong with the experimental setup, i.e., it would reduce uncertainty about (a) by excluding (b).[2] To establish ①, one has to show that I_X is a cause (simpliciter) of X. This is the case iff there is a variable set $\mathbf{V'}$ such that I_X is a contributing cause of X w.r.t. $\mathbf{V'}$. The latter is, according to interventionism, the case iff there exists a directed path from I_X to X in $\mathbf{V'}$ and an intervention variable I_{I_X} for I_X w.r.t. X such that changes induced on I_X's probability distribution by changes in I_{I_X} are associated with a change of X's probability distribution if the values of all off-path variables in $\mathbf{V'}$ are fixed by interventions. To establish that I_X is not a cause (simpliciter) of Y causing Y through X (condition ③), one has to show that there is no intervention variable I_{I_X} for I_X w.r.t. Y and no variable set $\mathbf{V'}$

[1] Recall that ② is dispensable.

[2] Recall that the intervention conditions ①, ③, and ④ guarantee correct causal inference.

(containing I_X, X, and Y) such that changes in I_{I_X} are associated with changes in Y's probability distribution if X is fixed by an intervention $I'_X = i'_X$. Note that establishing ③ is more challenging than establishing ① because a negative existential can only be established inductively. Finally, condition ④ holds iff I_X is independent of every cause (simpliciter) of Y that causes Y not through X. Also ④ can only be established on the basis of induction.

Summarizing, $q_{[I_X]}$ cannot be established with certainty on empirical grounds. Hence, Equation 1 does not allow us to infer that X is a cause of Y w.r.t. $\mathbf{V}$ (p) with certainty. This does, of course, not come unexpected. It follows from the fact that ③ and ④ can only be established inductively.[3] However, ③ and ④ can be confirmed by E' and E'', respectively:

E': No intervention on I_X w.r.t. Y we found so far is associated with changes in the probability distribution of Y if X's value is fixed by an intervention $I'_X = i'_X$.

E'': I_X is statistically independent of all causes (simpliciter) Z_i of Y we found so far that cause Y not through X.

Let E be the conjunction of E' and E''. According to the standard view that confirmation consists in probability increase, E confirms a hypothesis H iff $P(H|E) > P(H)$. Thus, E confirms $q_{[I_X]}$. If we assume that $r_{[I_X]}$ has already been established, then E, thus, also confirms $\exists I_X(q_{[I_X]} \wedge r_{[I_X]})$. With Equation 1 it then follows that E confirms p as well. If we now find that (ii) Y-instantiations brought about by interventions $I_X = i_x$ on X w.r.t. Y occur strictly before the corresponding X-instantiations, we can finally confirm the hypothesis that X is a backward cause of Y. The problem of finding evidence

[3]Note that identifying intervention variables within interventionism is always a partially inductive task regardless of whether the intervention variable is used for establishing forward or backward causal hypotheses.

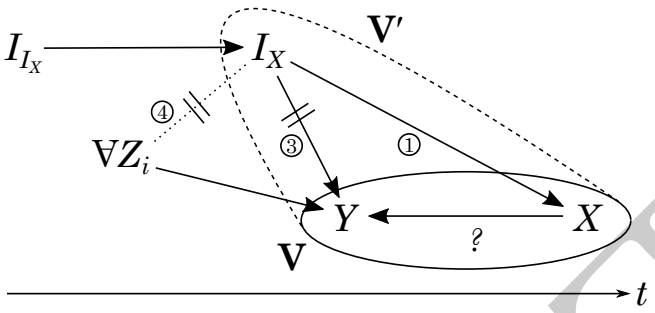

Figure 1: For X to be a backward cause of Y there must be an intervention I_X on X w.r.t. Y that induces Y-changes as demanded by **(M)** relative to some set $\mathbf{V}$. However, such a finding might cause doubt about whether I_X was actually suitable to infer a backward causal relation between X and Y; for the reasons mentioned at the beginning of section 4, (b) might be more likely than (a). Interventionism comes with the resources for testing whether I_X was suitable: Expand $\mathbf{V}$ to $\mathbf{V}'$ and test the ordinary causal hypotheses corresponding to conditions ①, ③, and ④. Evidence for them makes (b) less and (a) more likely.

for backward causation has been reduced to the task of confirming several ordinary causal forward hypotheses and checking for ordinary correlations between variables. Figure 1 illustrates and graphically summarizes the basic idea.

5 Comparison with other theories

Process causation. Process theories (e.g., Salmon, 1994) are based on the idea of causal processes and interactions, where a "causal process is a world line of an object that possesses a conserved quantity" (Dowe, 1995, p. 323) and a "causal interaction is an intersection of world lines that involves exchange of a conserved quantity" (ibid.). An event e_i is then considered to be cause of another event e_j iff e_i and e_j are connected by a causal process or a series of causal interactions. It is typically assumed that the cause occurs before the effect. To make sense of backward causation, one has to lift this requirement, which makes process causation a symmetric relation. As a consequence, it will fall out from

establishing that an event e_i is an ordinary forward cause of another event e_j that e_j is also a backward cause of e_i. Though causal processes and interactions are empirically identifiable, process causation clearly gives us way too much backward causation once one lifts the requirement that causes have to precede their effects. This problem is not shared by our interventionist approach: Though both analyses allow for establishing backward causal hypotheses by supporting forward causal hypotheses, our analysis in section 4 shows that allowing for backward causation within an interventionist framework does, contrary to the process analysis, not automatically render causal relations symmetric. Establishing backward causation empirically is a little bit trickier than just supporting the corresponding converse forward causal hypothesis empirically.

Laplacean causation. Laplacean causation (e.g., Hitchcock, 2012) is based on the idea that each possible state of a system can be described as a point (or vector) in a state space. The laws of mechanics rule how each possible state would evolve over time. Properties are identified with regions in a state space. The points within such a region represent all the possible states of a system in which the corresponding property would be instantiated. If each point within a region corresponding to a property P at an earlier point in time lawfully evolves into a larger region representing another property Q at a later point in time, then P is identified as a cause of Q.

How could Laplacean causation account for backward causation? First of all, note that the laws rule how a system evolves forward and backward in time. To allow for backward causation would then, again, consist in lifting the restriction that the cause must precede the effect. The empirical task would then consists in showing that any possible state instantiating a property P lawfully evolves into a state in the region corresponding to Q at an earlier time. The problem is that also this analysis would give us way too much backward causation. Assume

the domain of objects we are interested in are apples and that P stands for the property hanging on an apple tree and Q for the property being eaten. Now each point in the region corresponding to Q at any time can be evolved into a point within the region corresponding to P at an earlier point in time, simply because every apple being eaten grew on an apple tree. It would follow from Laplacean causation that being eaten is a backward cause of hanging on an apple tree, which seems quite absurd. It is easy to come up with thousands of similar examples using the same recipe. Similar problems do not arise for our interventionist analysis: Intervening on whether apples are eaten makes no difference for whether apples hung on apple trees before. Also note that finding out which points of a state space make up a region corresponding to a specific property might be practically unmanageable, while testing ordinary causal forward hypotheses is a quite straightforward empirical practice.

Counterfactual causation. According to counterfactual theories (e.g., Lewis, 1973), E_i is a cause of E_j if the following counterfactual holds: If E_i had not occurred, then E_j would not have occurred either. A counterfactual like this is true in the actual world $w_@$ if the closest possible world to $w_@$ in which E_i did not occur is a world in which also E_j did not occur. Lewis (1979) introduces the following similarity metric in order to flesh out the idea of closeness between possible worlds:

(S1) It is of the first importance to avoid big, widespread, diverse violations of law.

(S2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

(S3) It is of the third importance to avoid even small, localized simple violations of law.

(S4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

In principle, nothing in the definitions or the similarity metric above excludes that counterfactual theories can be used for analyzing backward causation. However, there are special conceptual problems with the similarity metric. (For details see Wasserman, 2015.) We add to these problems that counterfactual theories also fail in accounting for backward causation on empirical grounds. The evaluation of whether E_i is a backward cause of E_j requires an evaluation of the corresponding counterfactual in the closest possible world in which E_i did not occur. The problem is that it is unclear how one can evaluate the relevant counterfactual on empirical grounds. One might think that the similarity metric will identify the closest possible world in which E_i did not occur with one in which E_i was prevented by an experimental manipulation. Evaluating the relevant counterfactual would then simply amount to doing an experiment. The problem is that, as Woodward (2003, sec. 3.6) has argued, the similarity metric is far from being able to pick out the right world. What one would ultimately need to pick out the right world are criteria that guarantee that E_i has been prevented in this world via an experimental manipulation that satisfies the conditions specified in **(IV)**. (For details, see ibid.) But this means that for counterfactual theories to allow for establishing backward causal hypotheses on empirical grounds, one would have to replace Lewis's similarity metric with something that would come dangerously close to interventionism.

6 Conclusion

We developed an analysis of backward causation within an interventionist framework and argued that our approach does not fall victim to classical problems, which is an important minimal requirement for its adequacy. We then showed

how an interventionist treatment of backward causation can reduce the problematic task of directly testing for causal backward hypotheses to the quite harmless task of supporting several ordinary forward directed causal hypotheses. We thus provided a general methodology for establishing backward causation on empirical grounds. To further emphasize this advantage, we finally compared our approach to three other kinds of theories of causation.

Acknowledgements: We would like to thank Christian Loew and two referees for helpful comments.

References

Baumgartner, M., & Drouet, I. (2013, January). Identifying intervention variables. *European Journal for Philosophy of Science*, *3*(2), 183–205.

Black, M. (1956). Why cannot an effect precede its cause. *Analysis*, *16*(3), 49–58.

Dowe, P. (1995). Causality and conserved quantities: A reply to Salmon. *Philosophy of Science*, *62*, 321–333.

Faye, J. (2018). Backward causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Retrieved from https://plato.stanford.edu/archives/sum2018/entries/causation-backwards/

Hausman, D. (1998). *Causal asymmetries.* Cambridge: Cambridge University Press.

Hitchcock, C. (2012). Theories of causation and the causal exclusion argument. *Journal of Consciousness Studies*, *19*(5-6), 40–56.

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. (1976). The paradoxes of time travel. *American Philosophical Quarterly*, *13*(2), 145–152.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Mellor, D. H. (1991). Causation and the direction of time. *Erkenntnis*, *35*(1/3), 191–203.

Salmon, W. (1994). Causality without counterfactuals. *Philosophy of Science*, *61*, 297–312.

Wasserman, R. (2015). Lewis on backward causation. *Thought: A Journal of Philosophy*, *4*(3), 141–150.

Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.

Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research*, *77*(1), 193–212.