Getting Counterfactuals Right: The Perspective of the Causal Reasoner

Elena Popa

Abstract

This paper aims to bridge philosophical and psychological research on causation, counterfactual thought, and the problem of backtracking. Counterfactual approaches to causation such as that by Lewis have ruled out backtracking, while on prominent models of causal inference interventionist counterfactuals do not backtrack. However, on various formal models, certain backtracking counterfactuals end up being true, and psychological evidence shows that people do sometimes backtrack when answering counterfactual questions in causal contexts. On the basis of psychological research, I argue that while ordinarily both kinds of counterfactuals may be employed, non-backtracking counterfactuals are more easily used in causal inference because they are consistent with temporal order information embedded in the mental simulation heuristic, and they match reasoners' experience of causation. While this approach is incompatible with the ambitions of counterfactual theories that seek to establish the non-backtracking interpretation as the only legitimate one, it can provide support for perspectival views on causation and open further inquiry on the functions of causal and counterfactual thought in the context of causal models.

Introduction

The counterfactual theory of causation has been a central contribution to 20th century metaphysics. As the debate shifted from the ontological issue of what causation is to practice oriented questions, such as causal inference and the normative dimension of reasoning, counterfactuals continue to play a central role. Nevertheless, the question of backtracking arises in relation to both counterfactual theories of causation and accounts of causal inference based on counterfactuals. While the direction of causation is in line with the direction of time, counterfactuals may go both ways. Thus, a question arises regarding why only non-backtracking counterfactuals should be employed when accounting for causal dependence or making cause-to-effect inferences. This paper uses psychological evidence from causal reasoning to explain why non-backtracking counterfactuals are easier to grasp for causal reasoners and largely used in connection to causal inference. I will argue that the employment of mental simulation as a heuristic in causal thought involves non-backtracking counterfactuals due to its orientation from past to future. Furthermore, the experience of causation and connected counterfactuals is marked by this framework.

This article aims to bridge philosophical and psychological work on counterfactuals, explaining the intuitive appeal of non-backtracking counterfactuals, and their plausibility from a psychological perspective. As a contribution towards contemporary approaches to causation and counterfactuals, this paper highlights the importance of models that encompass uses of backtracking, and their integration with current models focusing mainly on prediction. Counterfactuals going from past to future are plausible given people's experience of time and causation, but other uses may allow for backtracking, as illustrated in the case of diagnostic reasoning. I suggest moving beyond the perspective going back to Lewis' account that non-backtracking counterfactuals are the right way of thinking about counterfactuals, while at the same time setting a more modest claim of their usefulness in psychological context. This can be used in a less metaphysically ambitious project of defining causality in relation to the perspective of the causal reasoner.

I start by discussing the philosophical background for counterfactuals and causal models highlighting the focus on the non-backtracking interpretation in philosophical approaches and describing formal models according to which backtracking counterfactuals can come out as true (section 2). I will then review counterfactual thought in both developmental context and adult causal reasoning (section 3). I subsequently discuss backtracking in psychological context, providing an explanation for the plausibility of the non-backtracking interpretation of counterfactuals in causal reasoning, and exploring further philosophical consequences in relation to causal projectivism (section 4).

2. Causation, counterfactuals, and backtracking

In this section I review counterfactual analyses of causation highlighting the issue of backtracking and its subsequent treatment in debates on counterfactuals, causation, and causal inference. From this starting point I will sketch out the main questions to be addressed in the paper: on how formal models can incorporate backtracking counterfactuals, on evidence regarding the use of backtracking in causal thought, on why it makes sense for causal reasoners to use non-backtracking counterfactuals, and further consequences for the philosophy of causation.

The paradigmatic analysis of causation through counterfactual dependence was introduced by Lewis (1974). Lewis's complete analysis of causation is beyond my purposes here, and I will focus on counterfactual dependence, which is sufficient (but not necessary) for causation on Lewis' account. The truth of the claim 'If A had not occurred, B would not have occurred' is sufficient for holding that A causes B. The truth values of counterfactuals are assessed through Lewis's possible world semantics, which is beyond my purposes here.[1] Backtracking is introduced as an objection in Lewis (1979), in relation to the project of analyzing both the direction of causation and the direction

[1] See Lewis (1973).

of time through counterfactual dependence. In short, the issue is that if one accepts that A causes B, one would accept the counterfactual (i) 'If A had not occurred, B would not have occurred'. Nevertheless, one could also accept the counterfactual (ii) 'If B had not occurred, A would not have occurred'. This would hold in cases when one would take the absence of the effect event to be indicative of the absence of the cause event. Lewis' resolution is to provide an account according to which only counterfactuals like (i) are true, and backtracking counterfactuals like (ii) are false: 'back-tracking arguments are mistaken: if the present were different the past would be the same, but the same past causes would fail somehow to cause the same present effects (Lewis 1979: 457). While subsequent debates have raised other objections resulting in an updated view by Lewis (2000), I will look instead at how counterfactuals are handled in the context of causal inference.

Woodward's (2003) interventionist approach to causation has a central counterfactual component. Again, without going into the complexities of Woodward's account, the upshot is that a variable A is the cause of another variable B within a variable set if an intervention variable changing the value of A would also change the value of B.[2] It should be pointed out that the interventions Woodward describes are not confined to actuality, but possibility. This brings counterfactuals into the picture: 'commitment to a manipulability theory leads unavoidably to the use of counterfactuals concerning what would happen under conditions that may involve violations of physical law' (Woodward 2003: 132). For a discussion of backtracking, further clarification is needed in connection to Lewis's account above. Woodward defines causation through a framework of causal models, thus involving causal concepts, and, unlike Lewis, does not aim for a non-circular account of causation. The definition of an intervention variable in relation to a variable set highlights the arrow-breaking feature: an intervention on a cause variable would leave previous variables intact (breaking the connection between variables instead of changing the values of previous variables). Woodward attacks Lewis' similarity criteria – which were intended to establish the falsity of backtracking counterfactuals – with an example where a backtracking counterfactual ends up true. In a complex scenario where a variable (C) generates several effects (E1...E5) at t1 and another effect (E*) at t2, intervening to see whether if E1…E5 had not happened E* would not have happened either would involve five small miracles preventing E1…E5, which is a problem for Lewis' claim that widespread miracles and violations of laws should be avoided. Thus, if a small miracle happens before C, that would involve a true backtracking counterfactual, namely that had E1…E5 not occurred, C would not have occurred (Woodward 2003: 139).[3] By contrast, Woodward's concept of intervention can handle this case without miracles or backtracking. Here,

[2] One issue is whether Woodward's account should be read as a theory of causation at all, see Popa (2015): section 1.2 for a review and arguments in this sense.

[3] See Glynn (2013) for a discussion of Woodward's counterexamples and a comparison to Lewis.

the similarity metric is rejected, while the interventionist apparatus accounts for similar types of counterfactuals to those discussed by Lewis being true in the context of causal inference. This is analogous to Pearl's approach that will be discussed below.

Taking a broad perspective where causation is connected to counterfactuals on ontological or epistemic grounds, it should be noted that for Lewis, as well as for cases when one reasons from causes to effects intervening on the cause to change the value of the effect variable within Woodward's account, backtracking counterfactuals end up false. A clarification to make here is that my aim is not to criticize causal models employing non-backtracking counterfactuals for the purposes of causal inference. Rather, I seek to highlight additional concerns arising from the use of a semantics of counterfactuals similar to that by Lewis. While Woodward focuses on cause-to-effect inferences and does not provide truth conditions for all counterfactuals, questions arise as to why this particular semantics for counterfactuals is appropriate in this context, and how it can be connected to semantics where backtracking counterfactuals end up true. My argument can supply an answer to the former question, namely that from the perspective of the causal reasoner non-backtracking counterfactuals are easier to grasp. In relation to this, it is also worth noting that there are philosophical approaches to causation that employ backtracking counterfactuals. For instance, Broadbent (2007, 2012) argues for an approach that accepts the truth of certain backtracking counterfactuals, thus handling counterexamples such as preemption or the problem of the transitivity of causation. Broadbent defends the truth of backtracking counterfactuals by rejecting Lewis' claim about counterfactual dependence being asymmetric, and highlighting that counterfactual reasoning can be useful when applied to the past (e.g., when tracing details about the origin of a certain event), and not only for predicting future effects (2012: 471-472). As my interest here lies in psychological aspects, I will not go into the details of this approach. However, one thing to point out is that Broadbent's arguments are consistent with both the formal approaches that accept uses of backtracking, as well as with the psychological findings I will review below regarding using counterfactuals in different inference patterns.

Having reviewed the philosophical background, and moving on to tracing the psychological plausibility of the non-backtracking interpretation of counterfactuals, there are two subquestions to clarify. Firstly, how can backtracking counterfactuals be made sense of formally? Secondly, are backtracking counterfactuals used when reasoning causally? And if so, what makes the use of non-backtracking counterfactual preferable? For the remainder of this section I will address the former subquestion, with the next two sections addressing the latter questions.

For my purposes here, I start by discussing the causal model for counterfactuals by Pearl (2009) stressing that even though Pearl rejects the similarity criteria, his view presupposes a semantics of counterfactuals similar to that of Lewis. The causal model approach by Pearl is thus better suited to address previous objections raised against Lewis' account (Starr 2019: 3.3). The point regarding similar semantics is made by several authors, though there is also work highlighting the differences. For instance, Skovgaard-Olsen et al. point out, 'Pearl showed that it was possible to derive the same conditional logics based on his structural semantics for counterfactuals as on Lewis' account' (2021: 75). Lassiter also describes Pearl's view as follows: 'theories of counterfactuals built around causal models have generally taken a stronger stance [than Lewis], ruling out backtracking as part of the definition of intervention' (2020: 13). Lassiter also notes other views as exceptions, such as that by Hiddleston (2005) and Lucas and Kemp (2015) which I will discuss below. Regarding divergences from Lewis' semantics, Briggs (2012) points out that there are differences both in truth conditions and regarding which inferences containing counterfactuals are valid. Briggs further argues that extending Pearl's model would yield into a different logic of counterfactuals. I will not explore these wider debates on semantics here, as my interest is in backtracking, and Briggs notably mentions that the causal models he discusses apply to non-backtracking counterfactuals (2012: 157).

Fisher (2017a) describes Pearl's semantics of counterfactuals as strictly interventionistic in contrast with that of Hiddleston (2005) and Fisher (2017b), raising an issue about handling backtracking. Fisher makes this point in a semantic context, highlighting that the epistemic purposes of cause-to-effect inference that Pearl focuses on may not necessarily lead to the best assessment of the truth values of counterfactuals in semantic context (2017b: footnote 22). I take this point to be important in spelling out the problem: Pearl's model can be viewed as answering a specific concern about causal inference and in this sense it should not be taken as an account about counterfactuals in general. Still, once the discussion moves to the question of assessing truth values for counterfactuals, the Pearl model has difficulty with uses that involve backtracking. This can be addressed by causal models that encompass Pearl's approach alongside interventions that allow backtracking, or by integrating the interventionist model with other semantics of counterfactuals. Before discussing such approaches, it is worth stressing that recent work by Pearl highlights three layers of causal inference: the first based on statistical associations captured by Bayesian networks which can incorporate diagnostic reasoning, predictive inference through intervention, and counterfactual inference (see Pearl and MacKenzie 2018). Thus, insofar as the causal models defended by Pearl are connected to Bayesian networks, ways of reasoning that involve prediction can be integrated with diagnosis. Regarding models integrating different semantics for counterfactuals, Schulz et al. (2019) use the critique raised in Fisher (2017a) as a starting point to show in experimental context that participants would accept the truth of backtracking counterfactuals such as ‘If the match lit, then if the match had not been struck it would not have lit’, which would come out as false under Pearl’s account.[4] Schulz et al. argue for an alternative notion of intervention, that allows backtracking. Particularly, the latter account integrates Pearl’s views within a broader approach to counterfactuals that includes backtracking.

Woodward’s account described above relies on causal models by Pearl. Particularly, the ‘graph surgery’ feature, where an intervention leaves previous variables in the system unchanged, is part of Pearl’s account. As mentioned above, although Lewis’ approach to counterfactuals relies on the possible worlds semantics, the result with regard to the falsity of backtracking counterfactuals is similar. Nevertheless, causal models involving counterfactuals can also take different forms. In a model by Hiddleston (2005) the main idea behind interventions is to leave the network intact (thus, not cutting the causal connections). This means that if a variable A is a cause of variable B, intervening on B would amount to changing the value of A, since the causal connection between the two is left intact. This would lead to the backtracking counterfactual ‘If B had not occurred, A would not have occurred’ being true (see Illustration 1). Rips (2010) compares the two models by Pearl and Hiddleston in experimental context. More recently, Lucas and Kemp (2015) introduced a model that builds upon both Pearl’s and Rips’s approaches, and allows backtracking. Khoo (2017) also defends a theory that allows both backtracking and non-backtracking counterfactuals.

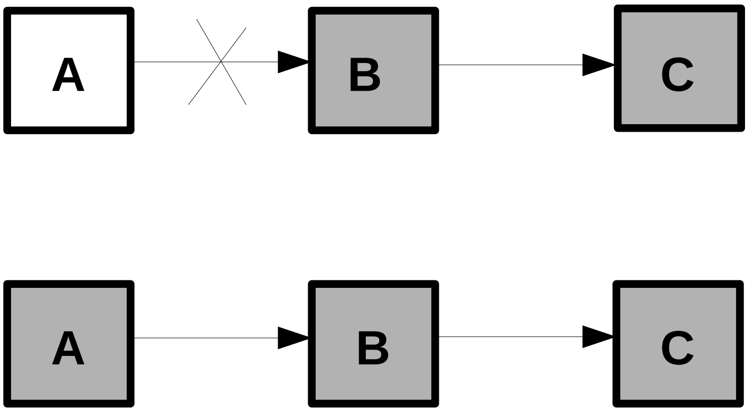

Illustration 1: The pruning and the minimal network approaches. For the former, if component B is not working (illustrated in gray), component A continues to work, but the causal connection between the two is cut. For the latter, if component B is not working, component A is also not working (they are both in gray). Thus, on the minimal network approach a backtracking counterfactual is true.

[4] On an interventionist reading, intervening to light the match would break the connection to its cause (striking it), thus the counterfactual ‘If the march lit, then if the match had not been struck, it would have lit’ would be true.

Given that formal models can accommodate both backtracking and non-backtracking counterfactuals for causal thought, the next question is an empirical one: whether backtracking counterfactuals are used in causal reasoning tasks. Before addressing this in the next section, one last thing to investigate is whether there are any particular strengths of the non-backtracking interpretation. One source for this is Woodward's (2014) functional account of causation, which ascribes his model a normative dimension: interventionist counterfactuals provide a framework that people *ought to* use when reasoning causally. Thus, on this view, one may claim that although both backtracking and non-backtracking interpretations are available formally, non-backtracking counterfactuals are the best suited for cause-to-effect inference. Woodward (2019) justifies normative models by an appeal to a means-ends relation: a feature of causal reasoning is assessed in relation to its successful employment in the pursuit of a goal. Woodward focuses on manipulability and control – a certain means of inferring causally would count as effective if it enables the successful control of relevant variables. Woodward (2019) further discusses how the normative connects with the descriptive in relation to work in psychology and experimental philosophy, highlighting the importance of empirical evidence. From this framework, interventionist counterfactuals have the advantage of ruling out confounders, which is an important part of causal knowledge. Still, the same functional perspective could also involve backtracking counterfactuals – tracing the cause of a particular effect can be used in future tasks involving control. In what follows, I will focus on descriptive aspects, namely how people *actually* reason about causality and what warrants the use of non-backtracking counterfactuals. Among other things, looking at time and the perspective of the causal reasoner will help explain the ease of using interventionist counterfactuals.

3. Causal reasoning and counterfactuals: psychological research

Discussing backtracking in the context of psychological research requires an investigation of counterfactual thinking from a broader perspective. I will first look into a debate on conditional and counterfactual reasoning in developmental context which will be relevant for my subsequent discussion of the normative approach mentioned above and related psychological accounts. I will then look into evidence of backtracking in adults' reasoning through causal models, highlighting that current research shows that people do not abide by one model only, or employ a single interpretation of counterfactuals.

Reviews of developmental evidence converge on conditional reasoning emerging around the age of 3, with counterfactual reasoning starting at 6 years of age (Gautam et al. 2019; Roese & Epstude 2017). However, whether younger children are able to answer counterfactual questions correctly has been subject to debate. Early work using stories told to children followed by questions about alternative turns of events suggested that counterfactual thought emerges between 3 and 5 years of age (Harris et al. 1996). These findings have been challenged, and explained through a distinction between conditional reasoning and counterfactual reasoning: younger children can answer questions about conditional scenarios correctly, while only children of 6 and above can answer the counterfactual questions (Rafetseder et al. 2010). This involves a more complex structure of counterfactual thought, namely creating an alternative scenario and integrating it with previous information (Rafetseder & Perner 2010). A debate between Weisberg and Gopnik (2013, 2015) and Beck (2015a, 2015b) focuses on whether counterfactuals are used earlier or later in development and whether counterfactual thinking is a continuous ability from early development to adulthood.

A more recent contrast is between work by McCormack et al. (2018) and Nyhout and Ganea (2019). The findings from the McCormack et al. (2018) study cast doubt over the claim that children under 6 can reason with counterfactuals. Given a scenario where a toy pig is toppled by mechanisms operated by disks of different colors, with one disk reaching it earlier, only the children over 6 were able to correctly say that had the first disk not been dropped, the pig would still have been knocked over by the second one.[5] This stands in contrast with findings by Nyhout and Ganea (2019), who use a device playing a tune when activated by blocks of certain colors to show that 4 and 5-year olds can answer counterfactual questions.[6] The point of contention here is whether children can reason with counterfactuals before the age of 6, and the answer is provided by their success at the task. One possible concern here is that the scenario from McCormack et al. (2018) may be viewed as too complex for children to grasp, and as such it may not shed light on their counterfacutal thinking. In response, I would like to point to further studies, like Rafetseder et al. (2013), using simpler scenarios that children should be able to grasp, but do not answer counterfactual questions correctly. The scenario by Rafetseder et al. (2013) involves a dwarf and a squirrel searching for nuts. The dwarf can pick up the nuts falling into a tree hut, taking them to the village, while the squirrel can pick up the nuts from the tree, taking them to the nest. Upon being presented one of the scenarios, say, the nut is in the tree and the squirrel is picking it up and correctly answering that the nuts will be in the nest, the children are presented with a counterfactual question: what if the dwarf had come to pick up the nut? Children below 6 would incorrectly say that the nut would be in the village.

Beck and Rafetseder (2019) explain the discrepancy between the results in the studies above, as well as the split between views attributing counterfactual thought to younger or older

[5] In the philosophical literature on causation this would count as a preemption scenario.
[6] This falls into the blicket detector paradigm (see Gopnik and Sobel 2000).

children through Hoerl and McCormack's (2019) dual systems account of temporal updating and temporal reasoning. For temporal reasoning, the child has to run a simulation of a past event, which is required in the McCormack et al. (2018) study. By contrast, in Nyhout and Ganea (2019) the screen still shows both blocks on top of the device and children have to only think of what would happen if one of the blocks were removed. Beck and Rafetseder (2019) take this to involve merely temporal updating, which is present earlier in development. While one may object that the preemption scenario may be too complex and thus not representative for assessing children's causal reasoning, the issue with the experiments in Nyhout and Ganea is that they do not require children to consider things having gone differently in the past.

Another potential objection here is that the Nyhout and Ganea experiments involve counterfactual reasoning in line with philosophical treatments of counterfactuals.[7] While the correspondence between the experimental work in psychology and philosophical work in counterfactuals is an interesting question that could further help clarify this debate, it is beyond my purposes here. I will sketch an answer that reasoning counterfactually in line with theories such as that by Lewis requires a set of advanced abilities, notably distinguishing between mere hypothetical reasoning and counterfactual thinking, and attributing these abilities to children below 6 should be backed up by further empirical work. Research on temporal reasoning such as the one mentioned above, for instance, suggests that relevant abilities emerge in older children.

While further research is needed in developmental context, it can be concluded that counterfactual reasoning emerges in childhood, with abilities easier to use by children emerging earlier in development. The point of contention regarding whether counterfactuals are used earlier rather than later and whether this is continuous with the use of counterfactuals in adulthood can be placed in the broader context of causal maps and the use of Bayes networks (for instance Gopnik et al. 2004; Schulz et al. 2007). Work on causal maps is consistent with normative approaches such as Woodward's (2014) above: children are said to follow certain rules for inference (including arrow-breaking interventions, or screening-off in probabilistic causation) which yield into correct judgments of causal structures, and also hold in adult causal cognition. Still, in the case of Bayesian models of causal learning, both normative and descriptive aspects are involved, and such views have been criticized for not clearly distinguishing between the two (Sloman & Fernbach 2008; Fernbach & Sloman 2011; Jones & Love 2011). As Fernbach and Sloman point out, 'violations of a model's predictions should be taken seriously and not explained away as due to the approximate way the optimal computation is implemented. And a rational analysis does not demonstrate

[7] I am grateful to an anonymous referee for bringing up this point.

rationality if people do not abide by it' (2011: 99). The authors further hold that more clarity about the descriptive and normative claims would render Bayesian approaches more falsifiable.

In relation to the problem of counterfactuals and causal reasoning, questions whether counterfactual thinking emerges earlier or later, and in accordance with a strictly interventionist model or a different one can be settled empirically if the descriptive claims are made clear. For instance, approaches such as Weisberg and Gopnik (2013) holding that counterfactual thinking is continuous throughout development and employing models using non-backtracking counterfactuals, do not explain how this relates to evidence of backtracking in adult reasoning. A set of questions emerges regarding where backtracking fits into this picture: is it simply an erroneous way of using counterfactuals? Is it a more sophisticated means of reasoning acquired later? Is it an ability that develops parallel to that of making cause-to-effect inferences? Conceptual contributions on the status of backtracking and causal reasoning such as the one sought by this article would help open the way for further empirical investigations in this sense. Particularly, the discussion in the following section would suggest a negative answer to the first question above: it sometimes makes sense to backtrack, although people generally employ non-backtracking counterfactuals when inferring causally.

By contrast with the developmental debate discussed above, the ability to answer counterfactual questions by adults is uncontroversial (Rafetseder et al. 2010, 2013). For instance, in the Rafetseder et al. (2010) study, adults had no problem answering counterfactual questions correctly. I will thus move on to the issue of backtracking. The study by Rips (2010) mentioned in section 2 explores the Pearl (2000) and Hiddleston (2005) models of counterfactuals in empirical setting, with the former excluding backtracking through the arrow-breaking feature, and the latter allowing it in order to keep the network intact (see Illustration 1). The participants were asked counterfactual questions about a system with different components, some of which are causally connected. In the case of a device where A causes B, the questions would be 'If component B were not operating, would A operate?' and 'If component A were not operating, would B operate?' (Rips 2010: 184). A positive answer to the first question would be indicative of backtracking. According to Rips' discussion, Pearl's model would predict that positive answers only are about the antecedent (e.g., if A causes B, intervening on B would leave A intact). According to Hiddleston's model, the answer should be negative (i.e., the network is left intact, and as such B not operating is an indication that A is also not operating). The results showed that people tend to backtrack, with similar results obtained in Rips and Edwards (2013). This runs in contrast with previous studies, such as Sloman and Lagnado (2005), that found a difference between observation and intervention: observation is associated with diagnostic reasoning (thus, involving backtracking), while intervention with causal reasoning (which falls in line with Pearl's model). It is important to point out that the questions and the task can be also interpreted in a different way: as respondents are only told that a component would not operate, and do not know whether the component was prevented from working by an intervention, the task may appear ambiguous, with people interpreting the workings of the system in different ways.[8] On this interpretation, Rips would not necessarily challenge the Pearl model. Still, the point may be raised in relation to semantics of counterfactuals that exclude backtracking (such as Lewis') and their overlaps with interventionist causal models: when faced with a choice between two interpretations, people do backtrack. Finding the conditions under which they do so is of further empirical interest and explored in the studies reviewed below.

Gerstenberg et al. (2013) ran a new set of experiments on similar devices as in Rips (2010) to shed further light on this. The significant finding is that people's answers tend to backtrack when asked about causes of the counterfactual state (thus falling in line with the Hiddleston's minimal network approach), but not when asked about the effect of the counterfactual state (thus falling in line with Pearl's pruning approach). Another finding is that people 'process counterfactual questions in a more local fashion rather than simultaneously considering the states of all variables in the system' (Gerstenberg et al. 2013: 2390). This will be relevant for my subsequent discussion on understanding counterfactuals and causality and on whether this understanding should be expected to fall in line with particular models, or whether people think of causes and effects without considering an entire variable system.

Han et al. (2014) address one shortcoming in the studies above, namely that talking about causality in the context of a device where components are represented as different variables may be too abstract a task to capture how people reason causally. Han et al. used questions in relation to various scenarios that would match the structure of the devices used in previous research (such as common cause or common effect). Examples of counterfactual questions include 'If John weren't drinking alcohol, then he wouldn't have brought a gift' and 'If John weren't drinking alcohol, then he wouldn't have acted wildly' (Han et al 2014: 2430). While in the case of the latter the causal connection between the antecedent and consequent is clear, the former makes sense for the causal reasoner only if there is another cause involved, for instance, receiving an invitation to the party would lead to both drinking alcohol and bringing a gift. The findings are interpreted by Han et al. as people backtracking just in case they make the counterfactual conditional true. This explanation runs in contrast with claims by Rips (2010), or Gerstenberg et al (2013) that link the use of different counterfactuals for explanation and inference, with the former tied to diagnostic reasoning and involving backtracking, and the latter following the order of causation that rules out backtracking.

[8] I am grateful to an anonymous referee for this point.

Unlike these previous explanations, Han et al. (2014) focus on the entire counterfactual and not only its antecedent, claiming that ‘backtracking counterfactuals can be considered a case of causal belief revision determined by the structure of the situation’ (2014: 2434). This is in line with earlier work by Dehghani et al. (2012) and Sloman and Walsh (2008). The contrast between the earlier claim by Gerstenberg et al. (2013) that people process counterfactual questions locally, and the reference to the structure of the situation by Han et al. (2014) should be noted. Still, the two claims are not completely incompatible, as the findings by Han et al. do not entail that people necessarily assume an entire network in accordance with various causal models, and neither do Gerstenberg et al. The issue at stake appears to be what kind of structure they are considering and how that determines whether they will resort to backtracking or not.

Having reviewed the psychological evidence on counterfactual reasoning and backtracking, one important finding to stress is that people’s judgments do not conform to solely one model, and this should be taken into account when discussing causal reasoning in relation to the respective models. Still, in the investigation of causal reasoning, especially in studies such as Rips (2010), Rips and Edwards (2013), Gerstenberg et al. (2013) that involve a distinction between explanation and inference, there is an assumption that at least in the context of reasoning from cause to effect, counterfactuals should not backtrack. This appears to echo earlier philosophical preoccupations of linking causation to non-backtracking counterfactuals reviewed in section 2. Given that formally both types of counterfactuals can be true according to different models, and that psychological evidence has shown that people do sometimes backtrack, the final question to ask is why non-backtracking counterfactuals tend to appear more natural in relation to reasoning causally. In the next section I provide an explanation of the preponderant use of non-backtracking counterfactuals in causal contexts on the basis of how people experience causality. In doing so, I will rely on further psychological work on mental simulation and causal thought.

4. Non-backtracking counterfactuals and the experience of causation

In will now provide an account of the usage of non-backtracking counterfactuals in causal thought. I employ two arguments in this sense: an argument from mental simulation and an argument from the experience of causality. The former holds that since mental simulation is often involved in causal reasoning, and it includes information on the direction of time, the counterfactuals that end up being connected to causation are those that follow the arrow of time. The latter will trace the use of non-backtracking counterfactuals to causal understanding and its connection to temporal succession. One consequence of both arguments is that the use of non-backtracking counterfactuals is not necessarily normative: mental simulation is used as a heuristic, and the intuitive understanding of causality does not provide sufficient conditions for causal inference. As such, the use of non-backtracking counterfactuals is not traced to their leading to better causal judgments as opposed to backtracking ones, but to their usefulness in the light of the workings of human cognition and its connection to the experience of time. Given that people can act to change the present or future and not the past, the forward-looking causal connections are the most relevant for the situation of the causal reasoner. Empirical work supporting these arguments includes developmental research on causal learning and the early connection between causation and temporal order, with the later connection to counterfactuals and their usage in adult causal reasoning, and the employment of mental simulation.

Before introducing the arguments, the connection between causation and counterfactuals previously discussed in relation to the philosophy of causation should also be traced in psychological investigations. While the studies reviewed above on counterfactual thought could help draw negative conclusions, for instance, regarding whether it makes sense to look for a connection between counterfactuals and causal learning in children under 6, the question whether people connect causation and counterfactuals is addressed in different research. One such relevant study was conducted by Gerstenberg et al. (2014), arguing that people's causal judgments are linked to counterfactual simulation. The participants were shown videos where two billiard balls would collide with the second ball either passing through a gate or being prevented from doing so. Both physical (a barrier) and non-physical (a teleport device) entities were involved in various scenarios. Participants were shown causal blocks and counterfactual blocks: the causal blocks showed the entire interaction while in the counterfactual blocks the video would stop at the time of collision. The authors interpret the results as follows:

> People make causal judgments by comparing what actually happened with what they think would have happened in the counterfactual world in which the causal event of interest hadn't taken place. They use their intuitive understanding of the domain in order to simulate what would have happened in the relevant counterfactual world (Gerstenberg et al. 2014: 526).

Thus, causal reasoning in adults is linked to counterfactuals understood according to the Pearl model. Related work includes a counterfactual model for causal reasoning in relation to physical events (Gerstenberg et al. 2015, 2020).

Having looked at empirical evidence regarding counterfactuals and causal cognition, I will now articulate the arguments. The argument from mental simulation holds that non-backtracking counterfactuals are employed in causal reasoning as part of the simulation heuristic. Given the wide use of this heuristic and particularly its connection to causal reasoning, non-backtracking counterfactuals are easier to employ for causal reasoners. Empirical evidence for this claim comes from psychological studies on the causal asymmetry. In a review by Lagnado and Sloman (2015) a section on the causal asymmetry concludes that ‘studies on the asymmetry of causal reasoning are consistent with the idea that people are much better able to run mental simulations forward from cause to effect than backward from effect to cause’ (3.16). Here, Lagnado and Sloman refer to Tversky and Kahneman’s (1974) earlier considerations on mental simulation, including aspects beyond probabilistic dependence, particularly spatio-temporal information.[9] While Lagnado and Sloman connect this to the geometrical-mechanical concept of causation, for the purposes of this paper I will leave the debate between defenders of these positions open.[10]

Mental simulation has been investigated in connection to various domains in psychology including decision making, self-regulation, memory, mental imagery, social cognition (Moulton & Kosslyn 2009: 1275). This wide range of uses helps illustrate the scope of causal thought: people simulate scenarios in order to make better choices, or to better navigate various social situations. Counterfactual thought and simulation are also discussed in the context of neuroscience by Van Hoeck et al.: ‘simulations provide the basis for constructing mental models of events and of imaging alternative realities ‘‘if only’’ different decisions were made or actions taken’ (2015: 2). One thing to stress here, though, is that mental simulation is a heuristic: a means of reasoning under conditions of uncertainty, which may yield correct judgments, but which may also be prone to error (Kahneman et al. 1982). In this sense, the contrast with approaches such as Bayesian models or causal maps should be noted. As causal maps rely on means of inferring causally that yield the correct causal structure, they would ascribe such methods to all causal reasoners. Still, as discussed above, such methods of inference rule out backtracking counterfactuals. The heuristics approach, and mental simulation in particular, presents a broader picture capturing the multitude of uses of counterfactuals. As simulation contains information on temporal order, that would explain the preference for non-backtracking counterfactuals, but backtracking may be used in cases where temporal order may not be as important. One consequence of adopting this picture is giving up the normative dimension that approaches such as Woodward (2014) have taken: people can and do reason about causation in multiple ways, and some ways are better adjusted to their situations than others.

The previous point brings me to the argument from the experience of causality: causal reasoners employ preponderantly non-backtracking counterfactuals because they match their

[9] Also see Kahneman (1995) for a discussion on counterfactuals specifically.

[10] See Waldmann and Mayrhofer (2016) for a discussion of different concepts of causation in psychological context.

understanding of causation, which involves the direction of time. The conceptual account of the asymmetry of causation I have defended elsewhere emphasizes the early connection between causality and time on the basis of psychological research on causal learning and causal reasoning (Popa 2020). On this model, causal reasoners link claims such as ‘A causes B’ to claims such as ‘A is temporally prior to B’. While this provides neither necessary nor sufficient conditions for causal inference - it leaves out simultaneous causation and time alone is a weak cue to causality - it explains how further cues to causality are shaped by this assumption, particularly the temporal asymmetry. Under this model, counterfactuals would be among those cues, and their use would presuppose an earlier understanding of causation in relation to time, and possibly other cues.

According to different studies, children rely on temporal succession in causal perception starting with the age of 3 or 4 (Bullock et al. 1982; Ranking & McCormack 2013). Regarding causal reasoning, 5 to 6 years old can infer causal structure on the basis of temporal cues (McCormack et al. 2014). Comparing this with the developmental evidence on counterfactual thought shows that children are able to perceive causality and temporal succession before they employ counterfactuals. Only evidence of the kind provided by Harris et al. (1996) could match this age range, but in the current state of scholarship no study has disentangled the children’s performance from their use of mere conditional reasoning. Regarding causal reasoning, if the age when children can think counterfactually is 6, as several of the studies cited above converge on, then the use of temporal cues also precedes it.

One potential objection here would come from research attributing counterfactual thought to children younger than 6. While this is currently an ongoing empirical debate, I would like to answer it from the perspective of existing scholarship. If the difference is explained by reliance on temporal updating, and temporal reasoning respectively (McCormack & Hoerl 2019), then the use of counterfactual simulation involves temporal reasoning, which in turn would involve an understanding of succession. As such, temporal reasoning would be one crucial step in the development of counterfactual thought even if one were to identify developmental precursors such as conditional reasoning.

Moving on to adult causal reasoning, Lagnado and Sloman (2004) show that adults are more successful at inferring causally when interventions are accompanied by temporal cues. This again highlights the connection between causality and time, which facilitates causal reasoning even when inferring through different cues. In sum, causality appears to be understood in temporal terms from early on, and once causal thought is shaped by the understanding of time, counterfactual reasoning assists in causal inference. Counterfactuals are employed from this frame because of the causal reasoner’s experience and understanding of time.

Having explained why non-backtracking counterfactuals appear to be the right kind of counterfactuals in the context of making causal judgments, I will use the remainder of this paper to place the proposed account in philosophical perspective. As mentioned above, the aspects of human cognition that render the employment of non-backtracking counterfactuals useful are not sufficient to assume this interpretation as correct on objective grounds. By ‘objective grounds’, I mean a set of truth conditions that make certain counterfactuals true, but not others. These would be required for defending counterfactual analyses of causation alongside causal realism. Earlier metaphysical attempts to define causation through counterfactual dependence tied to a particular interpretation of counterfactuals as non-backtracking may face this problem. Because non-backtracking counterfactuals work in the context of running mental simulations, they appear as more natural to use, but there is nothing beyond human cognition that makes it so. Nevertheless, the approach introduced here can be used in the metaphysics of causation by perspectival accounts: if causation is understood in a perspectival way, then the situation of the causal reasoner is in a certain sense constitutive of the concept of causation. Defending a counterfactual approach to causation on perspectival grounds would explain the use of non-backtracking counterfactuals with reference to the experience of time by the causal reasoner, with the arrow of time determining which counterfactuals are true. I will briefly illustrate this in relation to Price’s (2007) version of perspectivalism.

Price argues that the perspective of the decision maker is constitutive of causation. The architecture of deliberation is characterized by Fixtures and Options which can be Known or Knowable (2007: 275). Adding the agent’s temporal position to this, Price introduces the Fixed Past Principle, which he takes to be a part of naive physics: events that happened in the past are taken to be Fixtures and thus cannot be acted upon (2007: 277). This view, however, does not exclude the possibility of ‘an atemporal god, able to wiggle the material world in a much less temporally-constrained manner’ (2007: 280). This, however would do away with the concept of causation according to Price, since such intervention would change everything, making it impossible to single out particular causal connections.

Placing the main claims of this paper in the context of Price’s perspectivalism would shed more light on the causal reasoner’s temporal positioning and on how that relates to causal thought. Price does not discuss counterfactuals, but as long as counterfactuals may be used in causal thought they would fall under the same perspective. Thus, for the deliberation situation it is the counterfactuals that go from past to future that are relevant, since one can only act to change the future. Furthermore, the use of mental simulation in decision making, as discussed by Kahneman and Tversky (1981), would help sketch out the relevant psychological aspects of mental simulation and decision making. As this is used as a heuristic, however, the absence of an objective guarantee for its success should be emphasized.

This brings me to the question of normativity. As argued above, using non-backtracking counterfactuals can be justified by the causal reasoner's experience of causation and time, but this justification is subjective at best. Thus, the norms are adjusted to the causal reasoner's situation – if the counterfactuals are meant to be used for decision making, then following the arrow of time would be the most useful strategy. However, if reasoning takes place within a particular system where connections cannot be altered, then backtracking would make sense. Likewise, if the aim is diagnostic, then backtracking would work again. Thus, rather than relying on a semantics of counterfactuals that excludes backtracking as a norm, further specifications of where and when such methods are more effective are needed. This would also undermine a more ambitious project of expanding a non-backtracking interpretation of counterfactuals to all areas of causal thought.

6. Conclusion

This paper has explored investigations of counterfactuals, causation, and backtracking across philosophy and psychology. I have traced the focus on non-backtracking counterfactuals to philosophical approaches to causation and noted its subsequent use by normative and psychological models, arguing that there are no formal or empirical reasons to exclude backtracking. I have brought forward an explanation of the use of non-backtracking counterfactuals in the context of making a causal judgment through the mental simulation heuristic and people's experience of causation. This view can be of further use in providing projectivist approaches in philosophy with psychological support in explaining how the situation of the causal reasoner shapes causal concepts.

This approach helps move forward the debate on causation and counterfactuals by highlighting the need to shift the focus from justifying the employment of exclusively non-backtracking counterfactuals to exploring contexts where different types of counterfactuals are salient. At the same time, the proposed account helps shed further light on how people ordinarily use counterfactuals in causal contexts and decision making. This can inform future studies on the functions of causal and counterfactual thought, involving both conceptual and empirical work in developmental psychology and reasoning, among others.

References

Beck S. R. (2015a). Counterfactuals matter: A reply to Weisberg & Gopnik. *Cognitive Science* 40: 260–61.

Beck S. R. (2015b). Why what is counterfactual really matters: A response to Weisberg & Gopnik. *Cognitive Science* 40: 253–56.

Beck, S. R., & Rafetseder, E. (2019). Are counterfactuals in and about time?. *Behavioral and Brain Sciences*, 42, doi:10.1017/S0140525X18002157.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical studies*, *160*(1), 139-166.

Broadbent, A. (2007). Reversing the counterfactual analysis of causation. *International Journal of Philosophical Studies*, 15(2), 169-189.

Broadbent, A. (2012). Causes of causes. *Philosophical Studies*, 158(3), 457-476.

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, 27(1), 55-85.

Fernbach, P. M., & Sloman, S. A. (2011). Don't throw out the Bayes with the bathwater. *Behavioral and Brain Sciences*, *34*(4), 198.

Fisher, T. (2017b). Counterlegal dependence and causation’s arrows: causal models for backtrackers and counterlegals. *Synthese*, *194*(12), 4983-5003.

Fisher, T. (2017a). Causal counterfactuals are not interventionist counterfactuals. *Synthese*, *194*(12), 4935-4957.

Gautam, S., Suddendorf, T., Henry, J. D., & Redshaw, J. (2019). A taxonomy of mental time travel and counterfactual thought: Insights from cognitive development. *Behavioural brain research*, *374*, 112108.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36, No. 36).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *CogSci*.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2020). A counterfactual simulation model of causal judgment, URL =<https://psyarxiv.com/7zj94/>.

Glynn, L. (2013). Of miracles and interventions. *Erkenntnis*, *78*(1), 43-64.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, *71*(5), 1205-1222.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, D. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 1-31.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4), 843–887.

Han, J. H., Jimenez-Leal, W., & Sloman, S. (2014). Conditions for backtracking with counterfactual conditionals. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Harris, P. (2000) *The work of the imagination*, Oxford, England: Blackwell.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning, *Cognition* 61, 233–259.

Hiddleston, E. (2005). A Causal Theory of Counterfactuals, *Nous* 39(4), 632–657.

Hoerl C, McCormack T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. *Behavioral and Brain Sciences* 42, e244: 1–69. doi:10.1017/S0140525X18002157.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and brain sciences*, *34*(4), 169.

Kahneman, D. (1995). Varieties of counterfactual thinking. *What might have been: The social psychology of counterfactual thinking*, 375-395.

Kahneman, D., Slovic, S. P., & Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. Cambridge university press.

Kahneman, D., & Tversky, A. (1981). *The simulation heuristic* (No. TR-5). Stanford Univ CA Dept of Psychology.

Khoo, J. (2017). Backtracking counterfactuals revisited. *Mind*, *126*(503), 841-910.

Lassiter, D. (2020). Causation and probability in indicative and counterfactual conditionals, URL = http://web.stanford.edu/~danlass/Lassiter-causation-probability-conditionals-draft.pdf

Lewis, D. (1973). *Counterfactuals*. Blackwell Publishing.

Lewis, D. (1974). Causation. *The journal of philosophy*, *70*(17), 556-567.

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

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182-197.

Lucas, C.G. and Kemp, C., (2015). An improved probabilistic account of counterfactual reasoning. *Psychological review*, 122(4), p.700.

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The development of counterfactual reasoning about doubly-determined events. *Cognitive Development*, 45, 1-9.

Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1273-1280.

Nyhout, A. and Ganea, P.A., (2019). Mature counterfactual reasoning in 4-and 5-year-olds, *Cognition* 183: 57–66.

Pearl, J. (2009). *Causality*. Cambridge university press.

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

Popa, E. (2015). *Causation as Manipulability and Temporal Direction* (Doctoral dissertation, Central European University).

Popa, E. (2020). A Psychological Approach to Causal Understanding and the Temporal Asymmetry. *Review of Philosophy and Psychology*, 1-18.

Rafetseder, E., Cristi-Vargas, R., Perner, J. (2010) 'Counterfactual reasoning: developing a sense of "nearest possible world"', *Child Development* 81, 1: 376-389.

Rafetseder, E., & Perner, J. (2010). Is reasoning from counterfactual antecedents evidence for counterfactual reasoning? *Thinking and Reasoning*, 16, 131–155.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of experimental child psychology*, 114(3), 389-404.

Rips, L. (2010), 'Two Causal Theories of Counterfactual Conditionals', *Cognitive Science* 44: 2, 175-221.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107-1135.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In *Advances in experimental social psychology* (Vol. 56, pp. 1-79). Academic Press.

Schulz, L., Gopnik, A., Glymour, C. (2007) 'Preschool children learn about causal structure from conditional interventions', *Developmental Science* 10: 322-332.

Schulz, K., Smets, S., Velázquez-Quesada, F. R., & Xie, K. (2019, October). A logical and empirical study of right-nested counterfactuals. In *International Workshop on Logic, Rationality and Interaction* (pp. 259-272). Springer, Berlin, Heidelberg.

Skovgaard-Olsen, N., Stephan, S., & Waldmann, M. (2021). Conditionals and the hierarchy of causal queries. *Journal of Experimental Psychology: General*, *1*.

Sloman, S. A., & Fernbach, P. M. (2008). The value of rational analysis: An assessment of causal reasoning and learning. *The probabilistic mind: Prospects for Bayesian cognitive science*, 485-500.

Sloman, S. A., & Lagnado, D. A. (2005). Do We “do”? *Cognitive Science*, 29(1), 5-39.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology*, *66*, 223-247.

Starr, W. (2019). Counterfactuals, *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2019/entries/counterfactuals/>.

Walsh, C. R., & Sloman, S. A. (2007). Updating beliefs with causal models: Violations of screening off. MA Gluck, JR Anderson & SM Kosslyn, *A Festschrift for Gordon H. Bower*, 345-358.

Tversky A., Kahneman D. (1974) ‘Judgment under uncertainty: heuristics and biases’, *Science* 185:1124–31

Van Hoeck, N., Watson, P.D. and Barbey, A.K., (2015). Cognitive neuroscience of human counterfactual reasoning. *Frontiers in human neuroscience*, 9, p.420.

Waldmann, M. R., & Mayrhofer, R. (2016). Hybrid causal representations. In *Psychology of Learning and Motivation* (Vol. 65, pp. 85-127). Academic Press.

Weisberg D. S. & Gopnik A. (2013) Pretence, counterfactuals, and Bayesian causal models: Why what is not real really matters. *Cognitive Science* 37: 1368–81.

Weisberg D. S. & Gopnik A. (2015) Which counterfactuals matter? A response to Beck. *Cognitive Science* 40:257–59.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Woodward, J. (2014). A functional account of causation; or, a defense of the legitimacy of causal thinking by reference to the only standard that matters—Usefulness (as opposed to metaphysics or agreement with intuitive judgment). *Philosophy of Science*, *81*(5), 691-713.

Woodward, J. (2019). Causal Judgment: What Can Philosophy Learn from Experiment? What Can It Contribute to Experiment. *Advances in Experimental Philosophy of Science*, 205.