Removing the ‘Counterfactual’
from Counterfactual Theories of Explanation

Liron Karpati

5/15/22

Introduction

‘What is *explanation*?’ That is the question theories of explanation attempt to answer. Based on a 2018 collection of essays by prominent thinkers in the philosophy of explanation, the current “most promising… monist approach [to account for explanation] are counterfactual theories” (Reutlinger and Saasti 77). In this paper I aim to show why counterfactual theories of explanation are incorrect. They are incorrect, I claim, because there are examples of explanation in which counterfactual dependence does not figure into the account. First, I will make the question ‘what is explanation?’ precise by describing how the question is to be answered. Then I will lay out the main positions on the question, describing what constitutes a monist account of explanation. Additionally, I will put forth the representative counterfactual monist theory we will be evaluating, The Counterfactual Theory of Explanation (CTE). Finally, I will put forth arguments demonstrating why CTE does not account for certain non-causal explanations.

The Question

Before putting forth candidate theories to answer ‘what is explanation?’, we must first clarify what we seek when asking such a question. After asking people on the street this question you might hear responses getting at the sentiment ‘it’s when you give relevant information that clarifies something’ or ‘it answers *why?* questions.’ Indeed, we have a strong pre-theoretic intuition for when something is or is not an explanation such that when we see it, we usually know it. However, if we rely on intuition alone, we risk fooling ourselves into believing we have explained something when in fact we have not. Given that finding explanations is in large part the goal of science, having a clear way to discern explanations from non-explanations is desirable. What we seek is an analytic account of explanation.

We can begin with the following observation: explanation is a relation on the set of propositions. We talk about explanans E_1, E_2, …, E_n explaining C, the explanandum. For example, that ‘Suzy is sick’ is explained by ‘the water in the well is not clean and Suzy drank the water from the well.’ The well water not being clean and the fact that Suzy drank that water are the two explanans explaining why she got sick, the explanandum. Given that explanation is a relation, we can make our question more precise by asking for an account of explanation that specifies the necessary and sufficient conditions for when a set of explanans is in explanatory relation to an explanandum.

Perspectives On Explanation

In trying to specify the necessary and sufficient conditions that define explanation, there are three main perspectives in the literature: causal reductionism, pluralism, and monism. I will explain what these perspectives are and where they come from.

The question of precisely characterizing what constitutes an explanation arose primarily from philosophy of science. One recent success of philosophy of science has been to propose robust formalizations of the notion of causality, for example see (Pearl 2009) and (Woodward 2005). As causal inference became more prominent, it seemed as though much of what science does is ascertain causal relationships in the world. The success of causality in the philosophy of science has led science, in some places, to become synonymous with causal inference. We see this implicit in popular lectures like "Science Before Statistics: Causal Inference" (McElreath 2021). Given that the goal of science is to explain the world and that it does this very often by ascertaining causal factors, it is a plausible hypothesis that perhaps all explanations work by citing causal factors. This is the claim of so-called *causal reductionists.* They believe that all explanations are causal explanations, meaning they work by virtue of their reference to the network of causal interactions in the world. For example, a causal reductionist would highlight the fact that when we explained why Suzy got sick, we referenced the prior event that Suzy drank unclean water (which caused the sickness). It was by pointing to an event in Suzy's causal history that we explained why she got sick, a causal reductionist would claim.

Although many explanations are causal, it is too strong a claim to assert that all explanations are causal. There are many putative examples of non-causal explanation. Consider the following example, pointed out in (Lange 2013), that we will come back to later.

There was a famous question of whether the seven bridges of Konigsberg could be crossed exactly once, without going backwards, and so that you ended up where you started (call that a tour). The question was motivated by the fact that no one had yet found a tour. Euler explained *why that was the case; in fact he explained why it is impossible for there to be a tour. He first represented the bridges as a mathematical graph. Then he proved that one can only traverse the edges of a graph exactly once, without going backwards, so that you end up where you started (called an Eulerian cycle) when all the graph vertices have even degree (call this Euler's Cycle Theorem). The graph representing the Konigsberg bridges had a vertex of odd degree which proved there was no Eulerian cycle. Since the graph represented the bridges, there could be no tour of the bridges of Konigsberg.*

If we are to ask *'why can't one tour the Konigsberg bridges?'* the explanation would be Euler's Cycle Theorem.and that the graph representing Konigsberg has an odd degree vertex. Note how the explanandum, whether the Konigsberg bridges can be toured, is a naturalistic phenomena that was not explained by citing causal facts. It was explained by citing a mathematical fact.

Another example of non-causal explanation would be: there cannot be a perpetual motion machine because of the principle of energy conservation. We explain the impossibility of a particular naturalistic phenomena (a perpetual motion machine) by citing a non-causal fact (the principle of energy conservation). Such examples conclusively show that there are cases of non-causal explanation, so we reject causal reductionism.

Once we accept that there are non-causal explanations there are two theoretical positions we can take. Either we believe that there are different theories accounting for causal and non-causal explanation separately, *pluralism,* or we believe that there is one unified theory of explanation that covers both types, *monism.* While there is nothing wrong with pluralist accounts, per se, a single theory of explanation, by virtue of its generality, seems preferable to several different theories. We will restrict our attention to monist theories, specifically counterfactual monist theories.

The Counterfactual Theory of Explanation (CTE)

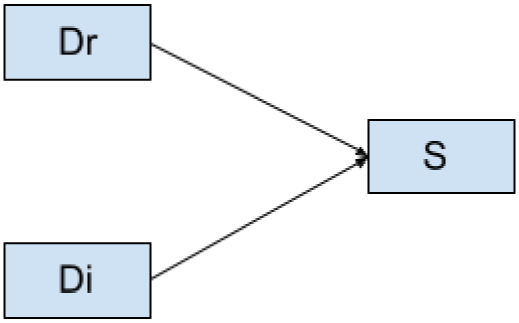
I take CTE to be the monist counterfactual theory outlined by Reutlinger in his essay "Extending the Counterfactual Theory of Explanation" (Reutlinger 2018). This will serve as a reasonable representative of counterfactual theories in general. As Reutlinger explains, "the core idea of the counterfactual theory… [is to analyze] explanatory relevance in terms of counterfactual dependence." He continues by characterizing CTE as the following *necessary* conditions:

1. Structure Condition: Explanation is a binary relation between a set of explanans and an explanandum, E. The explanans are further divided into three types. Nomic generalizations (G_1, …, G_m), initial conditions (IC_1, …, IC_n), and further background assumptions (A_1, …, A_p).
2. Veridicality Condition: The explanans and explanandum are approximately true.
3. Inference Condition: The nomic generalizations and initial conditions allow us to deductively infer the explanandum or the conditional probability $P(E \mid IC_1, \dots, IC_n)$.
4. Dependence Condition: "G_1, …, G_m support at least one counterfactual of the form: if the initial conditions IC_1, …, IC_n had been different than they actually are… then E, or the conditional probability of E, would have been different."
5. Minimality Condition: "No proper subset of the explanans … satisfies all of conditions 1 - 4."

The veridicality condition ensures that our explanation is indeed true. We can drop this condition because our goal is simply to understand what constitutes a well formed explanation. For example when a split brain patient is shown a command to get water to the side of the brain not responsible for linguistic production and then is asked why they went to get water, they often verbalize an explanation like "I was thirsty." This is a valid explanation, although it is not true. Reutlinger acknowledges this point; his aim with CTE was to characterize true explanations, not just valid ones. Our aim is to characterize valid ones so we drop veridicality.

To understand the rest of these conditions let us see how CTE accounts for causal explanation. We again consider the explanation Suzy is sick because she drank water from the well and the well water was dirty. To recall, our explanandum is the statement (i) Suzy is sick and our explanans are the statements (ii) Suzy drank from the well water and (iii) the well water is dirty. We will fit our explanans into the sub-categories of nomic generalization and initial condition soon. We ignore the veridicality condition for reasons already mentioned. When we try to check the inference condition we run into a slight issue. There is nothing clearly necessitating the deduction from our explanans to the explanandum. In a first order formalization we could have something like (i) $S(s)$ ‘Suzy is sick’, (ii) $Dr(s, w)$ ‘Suzy drank from the well water’, (ii) $Di(w)$ ‘the well water is dirty’. $Dr(s, w) \land Di(w)$ only entails $S(s)$ if we additionally believe $(Dr(s, w) \land Di(w)) \rightarrow S(s)$. But, $(Dr(s, w) \land Di(w)) \rightarrow S(s)$ is not one of our explanans.

What is actually going on is that when we say Suzy is sick because she drank water from the well and the well water was dirty, we are implicitly invoking a particular causal diagram in the background. We have three random variables S, Dr, and Di. S is 1 if Suzy is sick and 0 otherwise. Dr is 1 if Suzy drank from the well water and 0 otherwise. And Di is 1 if the well water is dirty and 0 otherwise. The causal diagram (let us assume a deterministic diagram for simplicity) that our explanation implicitly relies on looks like:



where $S = f(Dr, Di)$ is the structural equation governing how Dr and Di determine S.
$f(Dr = 1, Di = 1) = 1$ and for all other values of Dr and Di, f takes value 0. Now we can restate our explanandum and explanans more precisely as (i) $S = 1$, (ii) $Dr = 1$, (iii) $Di = 1$, (iv) we believe the above causal diagram to be a faithful model of the causal relationships in the world. Since $Dr = 1$, $Di = 1$, and $f(Dr = 1, Di = 1) = 1$ we can now properly derive that $S=1$. Therefore the inference condition holds. A non-deterministic causal diagram would allow us to infer the conditional probability mentioned as another case in the inference condition.

Technically, ‘Suzy is sick because she drank water from the well and the well water was dirty’ is not an explanation unless you have some implicit semantics accounting for how the explanans relate to the explanandum. We chose to make those background semantics explicit using a causal diagram to rigorously formalize the explanation. Our common parlance approximates this formal meaning but leaves the details implicit, similar to how when we say a curve is smooth we could say it more precisely as the curve is uniformly continuous.

Note how under our fully formalized explanation of the Suzy case we have clear nomic generalizations and initial conditions. The nomic generalization is the structural equation and the initial conditions are the states that the random variables Dr and Di take. Therefore the structure condition holds. We know minimality holds because without knowing that Di = 1 or Dr = 1, we cannot derive that S = 1. Also if we do not believe that the causal diagram (hence its structural equation) faithfully represents the real world's causal structure, then there is nothing mapping Di = 1 and Dr = 1 to S = 1.

Lastly, and crucially for CTE, we see that the dependence condition holds. The nomic generalization, our structural equation, supports the following counterfactual: Had Di = 0, we would have had S = 0. Thus we have that our nomic generalization supports a counterfactual of the desired form for the dependence condition. In fact, because causal explanations are those explanations that work by virtue of referencing the explanadum's causal history, we will always have the dependence condition satisfied in causal explanations, by following a similar accounting as above. While this does not show that the dependence condition is necessary for explanation *in general*, it is a reasonable hypothesis to think that it might generalize to non-causal explanations. This is the hypothesis CTE makes and it is the hypothesis I aim to show is incorrect.

On Counterfactual Dependence

As previously noted, under CTE, "causal and non-causal explanations are explanatory by virtue of exhibiting how the explanandum counterfactually depends on the explanans" (Reutlinger 74). The dependence condition is emphasized as the most important of the conditions. I will show with the following arguments that making counterfactual dependence a necessary condition for explanation impedes our ability to account for the Konigsberg example of non-causal explanation. First let us understand how CTE claims to satisfactorily account for the Konigsberg case. To summarize Reutlinger:

The explanandum is that it is impossible for anyone to tour the Konigsberg bridges. The explanans is Euler's Cycle Theorem (a nomic generalization) and "a statement about the *contingent* initial conditions that all parts [of Konigsberg] are actually connected to an odd number of bridges" (Reutlinger 84). Thus the structure condition is satisfied. We ignore veridicality as before, even though Reutlinger does not. The inference condition holds since Euler's Cycle Theorem and the initial condition entail the explanandum. Finally, the dependence condition is met since Euler's cycle theorem supports counterfactuals like "if all parts of Konigsberg had been connected to an even number of bridges, then people would not have failed to [tour] all the bridges" (Reutlinger 84). We clearly have minimality since from the nomic generalization or initial condition alone we can't get our explanandum. Italics on contingent were added for emphasis.

A Critique of Initial Conditions:

The first critique as to why CTE's accounting of the Konigsberg case fails, targets the assertion that the initial conditions are contingent. In asking for an explanation for why *the* bridges of Konigsberg cannot be toured, we mean those bridges with the particular structure to be found in Konigsberg. As soon as we entertain different initial conditions, different bridge structures, we are no longer talking about the same bridges of Konigsberg. Therefore the initial condition is not contingent. As a consequence, it does not make sense to entertain counterfactuals on the initial conditions, so the dependence condition is not satisfied despite the fact that Euler's explanation appears to be valid.

A related argument holds for the additional non-causal explanation: there can't be a perpetual motion machine because of the principle of energy conservation. There are no initial conditions to entertain counterfactual dependencies on since there are no contingent facts present in the explanans at all. Yet, the principle of energy conservation does seem to explain the fact that a perpetual motion machine is impossible.

In both these two arguments we see that an explanation need not contain initial conditions on which meaningful counterfactuals can be considered. This runs contrary to the dependence condition that CTE claims is necessary.

A Critique of Nomic Generalizations:

The second critique targets CTE's claim that nomic generalizations must support counterfactuals in which changing some initial condition(s) changes the explanandum. First, I claim that we must make Reutlinger's explanans more precise, very much like how we did for the causal explanation. This must be done because Euler's Cycle Theorem only has bearing on mathematical graphs; it does not say anything about what must be true given the relation of land masses in Konigsberg to its bridges. "That all parts [of Konigsberg] are actually connected to an odd number of bridges" together with Euler's Cycle Theorem does not entail anything, similar to how Suzy drinking from the well and the well water being dirty did not technically entail Suzy would get sick. We needed a mathematical model under which our explanans entailed our explanandum. I propose the following formalization.

"That all parts [of Konigsberg] are actually connected to an odd number of bridges" (Reutlinger 84) only matters if we accept that a particular graph, G, (in which parts of Konigsberg are to be formalized as nodes and bridges as edges) is a faithful representation of the Konigsberg bridge structure. By faithful, I mean an Eulerian cycle on G corresponds to a tour of the bridges. Euler's Cycle Theorem can tell us whether graph G has an Eulerian cycle. Then, because we believed that G is faithful to the Konigsberg bridges, the existence of an Eulerian cycle (or lack thereof) on G implies the existence of a tour on the Konigsberg bridges (or lack thereof).

Given this refined understanding, we can restate our explanans to be: Euler's Cycle Theorem and a statement that G, a particular graph, faithfully represents the structure of the Konigsberg bridges. This is exactly parallel to the causal explanation case with Suzy in which we refined our explanans to be more precise by including the claim that the causal diagram we chose faithfully represented the causal structure of the world.

Under the new formalization, it is not exactly clear what the initial conditions are. There are no contingent facts built into our model, like there was in the causal diagram. To stick with Reutlinger's account as much as possible, we might say that the initial condition is the graph G, since changing the graph can change the number of vertices with odd degree. I would argue that the graph G is not a contingent initial condition (just as the choice of causal diagram was not contingent in the causal case) because if we consider a different graph to be G, we are no longer talking about the representation we previously assumed to be faithful. However it could be argued that because our choice of representation determines G, it is contingent. Whether or not the initial condition, G, in this case is contingent or not will have no bearing on the argument. Let us assume that the graph G is a contingent initial condition, so that the critique in this section can work independently of the previous one. Let us see whether CTE can account for our more precise formalization (it will).

We can see that in the precise formalization the inference condition holds via the following proof. Since Euler's choice for his graph representing Konigsberg, G, has all odd degree vertices, by his Cycle Theorem there are no Eulerian cycles of G. By the faithfulness explanan, it follows that there are also no tours of the bridges of Konigsberg. Also since we are taking the graph G to be contingent, had there been a different graph with all even degree vertices, the Konigsberg bridges would have had a tour. Thus the dependence condition is satisfied. Minimality clearly holds as well. So, CTE does indeed account for the explanation. We proceed with the critique by considering a variation of this explanation, which I claim CTE does not account for.

Suppose I ask Euler in the 21st century if there is a tour of Konigsberg bridges. He again proceeds by representing the bridges as a particular graph, G*, assuming the graph is faithful. Now all that is left is to show that there are no Eulerian cycles of that graph. Let the graph G* have edges e1,..., en. Euler, now having a computer, does not bother to prove his general Euler's Cycle Theorem. Instead he simply enumerates the finitely many permutations of e1,..., en where each permutation gives a possible traversal order and then he checks if any permutation sequence exists in G* that leads to a cycle including all those edges. After seeing that none of the sequences work, he concludes that there are no Eulerian cycles. He writes a new theorem that G* has no Eulerian cycles, called Euler's Useless Theorem. By the faithfulness assumption he concludes that there is no tour of the Konigsberg bridges.

The computational enumeration is the proof of Euler's Useless Theorem just as Euler had some proof of Euler's Cycle Theorem. The proof itself does not figure into the explanation we care about. The explanans for our new explanation are Euler's Useless Theorem and a statement that G* faithfully represents the structure of the Konigsberg bridges. We consider two cases.

Case 1: If we do not consider Euler's Useless Theorem a nomic generalization then we have an explanation that works completely outside CTE's framing since it is not auxiliary and not an initial condition. Thus CTE would not actually specify necessary conditions as it claimed to. Case 2: If we do consider Euler's Useless Theorem a nomic generalization, it does not support counterfactuals of the initial conditions. Here is why. Euler's Useless Theorem talks about the specific graph Euler's computational proof considered, G*. If we consider counterfactuals about what would have happened had our choice of representation been G**, we cannot conclude anything since Euler's Useless Theorem only talks about G*. Therefore the dependence condition does not hold for our new example. One might try to avoid this result by claiming that when you change your representation to G**, Euler's Useless Theorem would then have been about G** too. But if this is the case then our nomic generalizations are changed by our initial conditions. This is contrary to how nomic generalizations are supposed to work.

In either case, CTE cannot account for the example explanation.

Conclusion:

In this paper we have considered a leading monist account of explanation, the Counterfactual Theory of Explanation. We found that although CTE nicely accounts for causal explanations, it cannot account for the Konigsberg non-causal explanation example or its variation. This was because CTE claims that the dependence condition is necessary and we showed that the dependence condition is not upheld in the Konigsberg case or its variation. Ultimately, we deem that CTE is too strong a theory because of the dependence condition.

If we remove the dependence condition (and veridicality condition as discussed) we are left with the structure condition, the inference condition, and the minimality condition as necessary conditions for explanation. Notably we can get rid of the distinction between explanan types in the structure condition because the distinction was only relevant to the dependence condition. The main relationship between explanans and explanandum in our reduced set of conditions is the entailment relation in the inference condition. I think we can get these conditions to sufficiently pick out what explanation is by supplementing it with an anti-symmetry condition. The anti-symmetry condition would say that an explanandum cannot entail the conjunction of its explanans. With this addition we would say that explanation *is* minimal, anti-symmetric entailment between a set of propositions (explanans) and an explanandum. Arguing for this account, The Entailment Theory of Explanation, will be the subject of another paper.

Works Cited

McElreath, Richard. “Science Before Statistics: Causal Inference.” *Youtube,* uploaded by Richard McElreath, 10 Sep. 2021, https://www.youtube.com/watch?v=KNPYUVmY3NM.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge, Cambridge University Press, 2009.

Reutlinger, Alexander. "Extending the Counterfactual Theory of Explanation." *Explanation Beyond Causation*, edited by Alexander Reutlinger, and Juha Saatsi, Oxford University Press, 2018, 74-95.

Woodward, James. *Making Things Happen: A theory of causal explanation*. Oxford, Oxford University Press, 2005.