Unraveling Sartorio's Difference-Making Principle

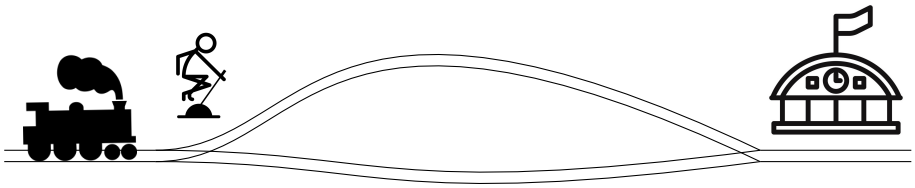
Dean McHugh, ILLC and Department of Philosophy, University of Amsterdam

Summary. We prove that a plausible but complicated principle concerning the meaning of *cause* – Carolina Sartorio's Difference-Making principle – is equivalent to a much simpler principle.

Keywords: Causation, difference-making, counterfactual dependence, modality, conditionals.

Abstract. In the philosophical and linguistic literature on causation, it is often said that for something to count as a cause it must 'make a difference' to the effect. As Lewis put it, "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it" (Lewis 1973:557). This difference-making component is illustrated in the following scenario, due to Hall (2000) and depicted below.

> An engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same. (Hall 2000:205)



Consider (1) in this context.

(1) a. The train reached the station because the engineer flipped the switch.
b. The engineer flipping the switch caused the train to reach the station.

The sentences in (1) are intuitively unacceptable. A plausible idea is that pulling the lever did not 'make a difference' to the train reaching the station.

But what does the difference-making requirement consist in? A compelling thought is that difference-making means counterfactual dependence: if the cause had not occurred, the effect would not have occurred. This correctly predicts (1) are unacceptable, for even if the engineer hadn't flipped the switch, train would have reached the station anyway.

However, as is well-known, the idea that causation requires counterfactual dependence is plagued by a host of counterexamples (see Lewis 2000, Hall and Paul 2003, Hall 2004, Halpern 2016, Beckers 2016, Andreas and Günther 2020 and many more). Here is an much-discussed example introduced by (Hall and Paul 2003:110) (the following formulation is from Hall 2004:235).

> Suzy and Billy, expert rock-throwers, are engaged in a competition to see who can shatter a target bottle first. They both pick up rocks and throw them at the bottle, but Suzy throws hers before Billy. Consequently Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's had not occurred, so the shattering is overdetermined.

Consider (2) in this context.

(2) a. The bottle broke because Suzy threw her rock at it.
b. Suzy throwing her rock at the bottle caused it to break.

Intuitively, the sentences in (2) are acceptable. (For experimental work confirming this see Walsh and Sloman 2011.)

The problem with analysing difference-making as counterfactual dependence is that (2) are acceptable even though the effect does not counterfactually depend on the cause: if Suzy had not thrown, the bottle would have broken anyway. This is quite a puzzle. The train track case shows that we need some notion of difference making. But if difference making is not counterfactual dependence, what is it?

Sartorio (2005) has a solution to this puzzle. About the train case, she writes:

> One thing that catches the eye ... is that, just as the *flip* doesn't make a difference to the [train reaching the station], the *failure to flip* wouldn't have made a difference to the [train reaching the station] either. In other words, *whether or not* I flip the switch makes no difference [to the train's arrival], it only helps to determine the route that the train takes [to the station]. (Sartorio 2005:74–75)

Sartorio distills this thought into the following principle.

> *Sartorio's Principle.* If C caused E, then, had C not occurred, the absence of C wouldn't have caused E.

Sartorio's principle represents a major breakthrough in our understanding of causal dependence, for it provides a principled way to distinguish preemption cases (such as the Billy and Suzy case) from switching cases. Let's first see what the principle says about the Billy and Suzy case. Suzy throwing her rock caused the bottle to break. What if Suzy hadn't thrown? In that case, it is clear that Billy throwing his rock would have caused the bottle to break. What about Suzy *not* throwing? Imagine if Suzy had not thrown (in that case Billy's rock would have hit the bottle and it would have broken anyway). Consider (3) in this context.

(3) a. Suzy not throwing her rock caused the bottle to break.
b. The bottle broke because Suzy did not throw her rock.

These are intuitively false. This is exactly what we need for (2) to satisfy Sartorio's Principle. Sartorio's Principle is therefore compatible with the truth of (2), as desired.

In contrast, imagine for the sake of argument that the engineer flipping the switch did cause the train to reach the station. As Sartorio (2005:75) points out, both flipping the switch and not flipping the switch make the same difference with respect to the train reaching the station (determining what route it took). So if the flipping the switch caused the train to reach the station, then for the same reasons, if the engineer had not flipped the switch, that would have also caused the train to reach the station. But this violates Sartorio's Principle, so the principle correctly predicts that the engineer flipping the switch did not cause the train to reach the station.

It therefore seems that Sartorio's Principle is something that we would like a semantic theory of *cause* or *because* to satisfy. But it is hard to see how to take any given semantics (one without a notion of difference making) and add Sartorio's principle to it. This is due to the principle's circular structure: it features *cause* in both the antecedent and consequent. We would like to find a way to 'unravel'

Sartorio's Principle – an equivalent principle with a non-circular structure. Our main contribution is to provide this principle, and prove its equivalence with Sartorio's principle.

Let us introduce some notation. For any sentences A, B and C, let $A[C/B]$ be the result of replacing every occurrence of B in A with C. For example, $\big((p \vee q) \wedge \neg q\big)[r/q] = (p \vee r) \wedge \neg r$. Now let $>$ denote the counterfactual conditional and consider the following principle.

> *The Perfection Principle.* For any sentences C and E, there is a sentence X such that C *cause* E entails $C > X$ and $\neg(C > X)[\neg C/C]$.

We call this the 'Perfection Principle' due to its similarity with an inference pattern known as conditional perfection (Geis and Zwicky 1971).

Our main result is that, under plausible assumptions about $>$, Sartorio's Principle is equivalent to the Perfection Principle. Those assumptions are given in (4), where $\diamondsuit\!\!\rightarrow$ is the dual of $>$, defined as usual by $A \diamondsuit\!\!\rightarrow C \equiv \neg(A > \neg C)$. The assumptions are all quite plausible.

(4)
a. **Nonempty domains.** $A > C$ entails $A \diamondsuit\!\!\rightarrow C$.
b. **Stability.** C *cause* E entails $C > (C$ *cause* $E)$.
c. **Idempotence.** $A \diamondsuit\!\!\rightarrow C$ entails $A > (A \diamondsuit\!\!\rightarrow C)$.
d. **Right weakening.** If C entails C' then $A > C$ entails $A > C'$.
e. If C *cause* E is true, then C is not a subsentence of E.

Theorem. *Sartorio's Principle is equivalent to the Perfection Principle, given the assumptions in (4).*

Proof. ($\Rightarrow$) Suppose Sartorio's Principle. Pick any sentences C and E and take $X = (C$ *cause* $E)$. Then by Stability, C *cause* E entails $C > X$. We also have the following chain of implications.

$$\begin{aligned}
C \text{ cause } E &\Rightarrow \neg C > \neg(\neg C \text{ cause } E) && \text{(Sartorio's Principle)}\\
&\Rightarrow \neg(\neg C > (\neg C \text{ cause } E)) && \text{(Nonempty domains)}\\
&\Rightarrow \neg(C > (C \text{ cause } E))[\neg C/C] && (C \text{ is not a subsentence of } E)\\
&\Rightarrow \neg(C > X)[\neg C/C] && (X = C \text{ cause } E)
\end{aligned}$$

Hence C *cause* E entails $C > X$ and $\neg(C > X)[\neg C/C]$. ($\Leftarrow$) Suppose the Perfection Principle. So $\neg C$ *cause* E entails $(C > X)[\neg C/C]$. Then by contraposition we have (†): $\neg(C > X)[\neg C/C]$ entails $\neg(\neg C$ *cause* $E)$. Observe the following chain of implications.

$$\begin{aligned}
C \text{ cause } E &\Rightarrow \neg(C > X)[\neg C/C] && \text{(Perfection Principle)}\\
&\Rightarrow \neg(\neg C > X[\neg C/C]) && \text{(Definition of } [\neg C/C])\\
&\Rightarrow \neg C \diamondsuit\!\!\rightarrow \neg X[\neg C/C] && \text{(Definition of } \diamondsuit\!\!\rightarrow)\\
&\Rightarrow \neg C > (\neg C \diamondsuit\!\!\rightarrow \neg X[\neg C/C]) && \text{(Idempotence)}\\
&\Rightarrow \neg C > \neg(\neg C > X[\neg C/C]) && \text{(Definition of } \diamondsuit\!\!\rightarrow)\\
&\Rightarrow \neg C > \neg(C > X)[\neg C/C] && \text{(Definition of } [\neg C/C])\\
&\Rightarrow \neg C > \neg(\neg C \text{ cause } E) && (\dagger \text{ and right weakening})
\end{aligned}$$

Hence C *cause* E entails $\neg C > \neg(\neg C$ *cause* $E)$, which is Sartorio's Principle. □

The key benefit of the Theorem is that the Perfection Principle is much easier to work with compared to Sartorio's Principle. To illustrate, suppose we have a theory of the meaning of *cause* that does not satisfy Sartorio's principle. The principle gives us a principled way to distinguish preemption cases (which do count as causes) from switching cases (which do not), so we would like to adjust our theory to satisfy Sartorio's principle.

The Theorem we have just proven gives us a way to do this. Plausibly, our theory says something about what would happen if the cause occurred ($C > X$), or something about what could happen if the cause did not occur ($\neg(C > X)[\neg C/C]$). This appears to be a minimal requirement, one that any theory of the meaning of *cause* would satisfy. If the former – our theory predicts that *C cause E* entails $C > X$ – we only have to add that *cause* entails that this no longer holds when we replace the cause with its negation: *C cause E* entails $\neg(C > X)[\neg C/C]$. And if the latter – our theory predicts that *C cause E* entails $\neg(C > X)[\neg C/C]$ – then as before, we only have to add that *cause* entails that this no longer holds when we replace the $\neg C$ with C: *C cause E* entails $C > X$. The resulting theory satisfies the perfection principle, so our Theorem guarantees that it also automatically satisfies Sartorio's principle.

C cause E just in case *C proto-cause E* $\wedge$??

To take an example from the literature, Beckers and Vennekens (2018) propose that *cause* requires that, if the cause had not occurred, the absence of the cause would not have produced the effect: $\neg C > \neg(\neg C$ *produce* $E)$. To make their account satisfy Perfection Principle, we need only add a conjunct stating that this does not hold when we we replace the cause with its negation: $\neg(C > \neg(C$ *produce* $E))$. Our Theorem guarantees that the resulting account automatically satisfies Sartorio's Principle.

References

Andreas, Holger and Mario Günther (2020). "Causation in terms of production". *Philosophical Studies* 177.6, pp. 1565–1591. DOI: 10.1007/s11098-019-01275-3.

Beckers, Sander (2016). "Actual Causation: Definitions and Principles". PhD thesis. KU Leuven. URL: https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1656621&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US.

Beckers, Sander and Joost Vennekens (2018). "A principled approach to defining actual causation". *Synthese* 195.2, pp. 835–862. DOI: 10.1007/s11229-016-1247-1.

Geis, Michael L. and Arnold M. Zwicky (1971). "On invited inferences". *Linguistic inquiry* 2.4, pp. 561–566. URL: www.jstor.org/stable/4177664.

Hall, Ned (2000). "Causation and the Price of Transitivity". *Journal of Philosophy* 97.4, pp. 198–222. DOI: 10.2307/2678390.

— (2004). "Two concepts of causation". *Causation and counterfactuals*. Ed. by John Collins, Ned Hall, and Paul Laurie. MIT Press, pp. 225–276.

Hall, Ned and Laurie A. Paul (2003). "Causation and preemption". *Philosophy of Science Today*, pp. 100–130.

Halpern, Joseph Y (2016). *Actual Causality*. MIT Press.

Lewis, David (1973). "Causation". *Journal of Philosophy* 70.17, pp. 556–567. DOI: 10.2307/2025310.

Lewis, David (2000). "Causation as Influence". *Journal of Philosophy* 97.4, pp. 182–197. DOI: 10.2307/2678389.

Sartorio, Carolina (2005). "Causes As Difference-Makers". *Philosophical Studies* 123.1, pp. 71–96. DOI: 10.1007/s11098-004-5217-y.

Walsh, Clare R. and Steven A. Sloman (2011). "The Meaning of Cause and Prevent: The Role of Causal Mechanism". *Mind & Language* 26.1, pp. 21–52. DOI: 10.1111/j.1468-0017.2010.01409.x.