Irreversible (One-hit) and Reversible (Sustaining) Causation

Abstract: This paper explores a distinction among causal relationships that has yet to receive attention in the philosophical literature, namely, whether causal relationships are reversible or irreversible. We provide an analysis of this distinction and show how it has important implications for causal inference and modeling. This work also clarifies how various familiar puzzles involving preemption and over-determination play out differently depending on whether the causation involved is reversible.

1 Introduction.

Recent work in philosophy of causation has explored differences *among* causal relationships, rather than just describing the contrast between relations that are or are not causal. This work assumes that relationships that are causal in some broad sense (e.g., in virtue of satisfying an interventionist conception of causation) can differ among themselves in ways that deserve philosophical attention. For example, causal relationships can differ in their degree of stability, specificity, proportionality, and their "speed" or temporal structure (Woodward 2010; Blanchard et al. 2018; Ross 2018). This paper explores an additional distinction among causal relationships, which to the best of our knowledge has not yet been discussed in the philosophical literature–namely, whether the relationship is reversible or irreversible.[1] By a reversible causal relationship between X and Y we mean a relationship such that if a change dx in X causes a change dy in Y at time t, then a reversal of this change in X at a later time can reverse or "undo" this effect. An irreversible causal relationship is a causal relationship that is not reversible. Slightly more formally: A causal relation between X and Y is irreversible if, when a change from $X= x_1$ to $X= x_2$ in X causes a change in Y from $Y= y_1$ to $Y=y_2$ at time t, it is not possible to change $Y=y_2$ to $Y= y_1$ by changing $X= x_2$ to $X= x_1$ at any later time $t'> t$. For example, if turning a light switch on causes a light to shine, and turning it off "reverses" this, extinguishing the light, this causal relation is reversible. Similarly, if the close proximity of the moon causes high tides, and the moon's different position at a later time reverses this, causing low tides, we have reversibility. As we note below, this sort of reversibility is commonly assumed in causal modeling frameworks and it is characteristic of most fundamental physical laws. It is striking, however, that it is *not* present in many of the most prominent cause-effect examples in the philosophical literature. When Suzy throws a rock that breaks a bottle, her rock cannot be "unthrown" in a way that restores the broken bottle. Similarly, if a sniper successfully shoots the enemy or the King is poisoned, neither of these effects can be undone by restoring the cause to its earlier state. These "one-hit" causes (as we shall call them) are "irreversible" with respect to the effects they produce. As we shall see this has important implications for how we should model such cases—we should not model them by means of a framework that assumes reversible causation. It is also the case that various familiar puzzles involving preemption and over-determination play out differently depending on whether the causation involved is reversible.

[1] The one discussion that we know of in the social science literature is Lieberson (1985).

Irreversible causation is not just important in everyday cases–-it figures in many areas of science including, perhaps especially, the social, behavioral, and biological sciences. Lieberson (1985) discusses the establishment of English as the dominant language for international communication as one example. While many causal factors led to this, including the geopolitical influence of the U.K. and the U.S., diminishing this influence seems unlikely to remove the dominance of English, at least in the short run. Once a language is widely adopted, the large costs of moving to an alternative language may keep the former in use even if the causes of its initial adoption are removed. As another illustration, based loosely on Wilson (1987) and Kearney and Wilson (2018), suppose that in an area in which wages and marriage rates are previously high, wage rates decline, which causes a decline in marriage rates (due to fewer eligible partners to support families). If wage rates subsequently increase, we should not necessarily expect an increase in marriage rates—the prior decrease in marriage may involve changes in attitudes and values that persist and continue to lead to low marriage rates even if wages increase. As Lieberson puts it, once an effect occurs it can "create circumstances that will perpetuate itself even if the initial causal variable is reversed" (1985, p.76). As these examples suggest, irreversible causation is often present in the social realm because of changes in "culture", expectations, and memories that remain when their external causes do not persist. Other cases of irreversible causation come from the biomedical sciences. For example, some causes damage biological structures without providing a means to reverse this damage. If five years of heavy smoking causes lung cancer, five subsequent years of smoking cessation will not reverse this. In other cases, irreversible causation may be part of a normal developmental process as when a pluripotent stem cell acquires the characteristics of a specific tissue type.

2. A More Detailed Look at Irreversibility

We have defined irreversible causal relations as those in which a change in *C* causes a change in *E* but returning *C* to its original state does not return of *E* to its original state. This covers two distinct possibilities. In some cases (a) reversal of the effect is either strictly impossible as when a patient dies or it is understood to be impossible for all practical purposes--when a letter is burned there is no practical possibility of collecting the combustion products to reconstruct the letter[2]. In other cases, (b) the effect can be reversed, but only through factors other than the initial triggering cause. For example, when a falling bookshelf causes someone to break a bone, returning the bookshelf to its original position won't unbreak the bone, but physiological processes may be in place to restore the bone back to an unbroken state. Similarly, it might be possible in principle to restore a broken bottle to its unbroken state but this will involve operations on factors other than the cause of the breaking. We treat both (a and b) as cases of irreversible causation.

The notion of reversibility just described should be distinguished from the feature of time reversal invariance possessed by many fundamental physical laws. The latter has to do with whether, when a process is permitted by a law, the process described by its time reversal (the result of substituting -t for t and perhaps making substitutions for other quantities in the law, such as the replacement of the magnetic field B with -B) is also permitted. Among many

[2] It may also be that the reversal of the cause is impossible or ill-defined as when a rock cannot be "unthrown".

other differences, our notion of reversibility is not characterized in terms of operations on a time variable or its derivatives, although it does involve relationships between values of variables at different times-- see above. It is nonetheless the case that many fundamental physical laws describe relationships that are also reversible in our sense. For example, when current is passed through a wire, an electromagnetic field is created and when the current is stopped, this field will disappear. If the distance *d* between two masses is increased to 2*d,* causing a decrease in the gravitational force between them, returning the distance back to *d*, causes the gravitational force to return to its original value, again illustrating reversibility in our sense. In these cases, the present value of some variable in a system just depends (for all values of this variable) on the present value of other variables characterizing the system—when this feature is present, the causal relation in question will automatically be reversible in our sense. This feature is not present in systems exhibiting hysteresis, in which the present state of some variable depends not just on the present state of other variables characterizing the system but on the history of the system—the causes to which it was exposed in the past and perhaps the temporal order in which these occur.

However, our notion of irreversibility is stronger than the notion of a system exhibiting hysteresis. In the latter case, it might be possible to undo an effect by appropriately reversing the various causes in its history. By contrast in irreversible causation, one cannot undo the effect by undoing its original causes. Irreversible causation is thus stronger than mere influence of the past; it involves past causes putting a system in a state which cannot be further influenced by presently restoring those causes to their original state or perhaps cannot be further relevantly changed by operations on any present causes. It may well be true that at a fundamental physical level, all physical laws are reversible in our sense, with apparent irreversibility reflecting the fact that we are operating with models and representations that omit relevant variables. But in common sense contexts and in much of science, including biology and the social and behavioral sciences, we are stuck with non-fundamental theories and thus have to deal with phenomena like irreversibility and hysteresis.[3]

We seem to have different mental models when we think about reversible and irreversible causation and we think in terms of different paradigmatic applications. To begin with the latter, the cases in which irreversible causation is operative virtually automatically call

[3] An interesting question, which we will not try to explore in detail, has to do with the relationship between irreversible causation and the second law of thermodynamics. Many of our examples of irreversible causation involve what one informally thinks of as entropy increase – the king who transitions from live to dead, the broken bottle etc. In those cases of irreversible causation in which the effect is reversible (but not by reversing the cause) the causes that might be employed to reverse the effects typically involve the very precise coordination of a number of factors in a way that looks (locally) anti-entropic: someone needs to piece together the broken glass and so on. In such cases, if the effect is reversible at all this usually seems to require something more complicated and organized than the mere reversal of the original cause. On the other hand, it is not clear, how if at all, the notion of entropy might be applied to some of our examples like falling marriage rates—this is why we say “many” above.

for “actual cause judgments” about particular causal interactions. This is to be expected—if a change in the cause puts the effect in a state which cannot be reversed by reversing the cause, we are naturally led to treat the case in terms of a “one-off” judgment that focuses only on that particular interaction, since there is no possibility of the cause “doing” anything more to the effect at other times. The one-off change in marriage rates in the example above illustrates this. This is not to say that one cannot make actual cause judgments about examples involving reversible causation—when a light is caused to go on or off depending on the state of a switch (reversible causation), turning the switch from off to on at some particular time is regarded as the actual cause of the light going on. However, in this case, unlike examples involving irreversible causation, one also thinks about the system in terms of a type-level causal relation that is repeatable for that system. It is thus unsurprising that the philosophical literature on judgments of actual causation tends to focus almost entirely on examples involving irreversible causation—these are cases that naturally invite such judgments.

When an interaction involves irreversible causation our mental picture is that the cause “acts” just once to produce the effect, but once the effect has occurred, the cause does not continue to act to keep the effect in place—the impact of the rock shatters the bottle but once this happens, the rock and its impact don’t continue to operate to keep the bottle shattered. Both the cause and the effect involve changes or transitions from one state to another (no impact to impact, bottle unshattered to shattered) that take typically place over identifiable and often short time intervals—thus an interaction that is naturally coded as a relation between “events”. By contrast, in typical cases of reversible causation, the cause continues to act as long (but only as long) as the effect is present—it “sustains” the effect. For example, the weight that extends the spring is naturally viewed as an ongoing cause rather than a discrete event.

3. How the distinction between reversible and irreversible causation matters for both inference and modeling.

Standard causal modeling frameworks typically assume that causal relations are reversible. For example, when one writes down an ordinary linear regression equation

(3.1) $Y= a_1 X_1+\ldots+ a_n X_n$

this implies that if, say, X_1 is increased by dX_1, Y will increase by $d\, a_1 X_1$ and that if X_1 is then decreased by the same amount, Y will return to its original value—the equation implies that the value of Y just depends on the current value of X_1 etc. and not on the previous values of this variable. In a context in which different values of $X_1\ldots X_n$ are observed for a system over time, a mistaken assumption of reversibility can easily lead to incorrect inferences. Suppose, to return to an earlier example, we observe marriage rates and wages in a particular area over time, and we attempt to represent the relation between these two variables via a reversible model like (3.1). Suppose, as assumed above, the actual causal relation is irreversible: if one begins with a state in which wages and marriage rates are high, a first time decline in wages will cause a decline in marriage but once the decline in marriage occurs, it becomes “locked in” (due to changes in attitudes and values) and subsequent increases in wages will not cause increases in marriage. Observing a time series of wage and marriage levels over time, with the former but not the latter changing over time, one will observe virtually no (or a very weak) correlation between these two variables, which may lead to the mistaken conclusion that there was never any causal relation between wages and marriage rates. But, ex hypothesi, there was such a relation for the initial change in wages although there is no subsequent relation between wages and marriage at later times. We need to use a model other than (3.1) which is sensitive to the possibility of irreversible causation to detect this possibility.

Consider next models of actual causation using structural equations—a currently very popular enterprise (see, e.g Halpern (2016)). Begin with a very simple possibility: Suzy throws a rock, it hits an *intact* bottle and the bottle shatters, with the causal relations assumed to be deterministic. A standard way of representing this -- again see Halpern-- is with binary variables *ST* (for Suzy throws), *SH* (representing whether Suzy's rock hits an intact bottle) and *BS* (for bottle shattering). *ST=1* if Suzy throws, *ST=0* if she does not and *SH* and *BS* similarly take values {1,0}.

The usual assumption is that the accompanying equations are:

(3.2) *SH=ST*

(3.3) *BS= SH*,

with *ST=1*

from which, on standard accounts of actual causation, one concludes that *ST=1* causes *BS=1*, in accord with intuitive judgment[4].

Note, though, that this model represents the causation involved as reversible or at least fails to represent that the causation involved is not reversible. This is reflected in the fact that (3.2) – (3.3) make no reference to time -- such reference is required to explicitly model irreversibility. Suppose that we incorporate such references -- we interpret *ST* so that it can take different values over time, writing *ST(t)=1* to mean Suzy is throwing at time *t* and similarly for the other variables. Then, if *ST(t)= 0* at any time *t*, we do have *SH (t) = 0* in accord with (3.2). On the other hand, if Suzy throws for the first time at *t*, then if she throws again at some later time, so that *ST(t+d)* also =1, it will not be the case that *SH (t+d)* =1 contrary to what a time-indexed analog of (3.2) implies. Moreover, once Suzy throws the subsequent state of *ST* will have no influence on the *BS* variable, again contrary to what (3.2) and (3.3) imply. All of this is a reflection of the fact that the causation involved is irreversible—it is part of our understanding of the problem that once *SH=1,* this event cannot undone by *ST= 0* at some later time and similarly for *BS=1*.

It thus seems that when we employ equations like (3.2)- (3.3) we tacitly understand them as relying on additional constraints or interpretive requirements that are not made explicit either in this representation or a simple time-indexed variant. In particular there are constraints *among* the values the variables can take at different times: e.g., if *SH(t) =1*, *SH (t')=0* for all times *t' > t.* There is also the implicit constraint that the bottle is intact before Suzy

[4] Halpern (34, 2016) also considers a considers a model of this example in which the variables are time-indexed.

throws for the first time—if it is not, neither *SH* or *BS* will depend on *ST*. A similar constraint should be imposed on *BS(t):* once *BS(t)=1*, *BS(t) =1* at all later times.

When dealing with cases containing the simple structure just described these subtleties may seem not to matter much since, as noted above, such cases involve one-off, actual cause judgments in which our focus is just on the causal relationships at the single time when Suzy's rock strikes the bottle and not at any other time. Note though that even in this case a fully explicit accurate modeling seems to require reference not just to time but to relations among the possible states the variables can assume at different times.

The role of considerations having to do with reversibility becomes more salient when we look at more complex cases involving actual cause judgment. Consider the following model for late pre-emption example that Halpern discusses in his (2016)[5]. Both Billy and Suzy throw rocks at the bottle. Suzy's rock hits the bottle and it shatters but if Suzy's rock had not hit the bottle, Billy's rock would have hit it a moment later and it would have shattered. As it is, Billy's rock just passes through the empty space where the bottle had been. Our clear judgment is that the impact of Suzy's rock caused the shattering. Halpern models this as follows (with *ST=1*)

(3.4) *BS= SH or BH*

(3.5) *SH= ST*

(3.6) *BH= BT and not SH*

What is noteworthy here is equation 3.6 and the arrow from *SH* (Suzy hits intact bottle) to *BH* (Billy hits intact bottle). Thus, *SH* is represented as *causing BH*—Halpern says that *SH=1* "prevents" *BH=1* (and presumably causes *BH=0*). As Halpern acknowledges this *SH*→*BH* relation is required for the model to reproduce the judgment that *ST* and not *BT* causes *BS=1.*

This particular modeling choice seems problematic for several reasons, which trace back to the distinctive features of irreversible causation. As a warm-up observation, note that there is a natural sense of "prevent" in which *SH =1* does not seem to prevent *BH =1.* According to this sense *X=1* prevents an outcome *Y=1* by interfering or interacting with some other cause *Z* of *Y* (a *Z* that "threatens" to cause *Y*) that would have caused *Y=1* if *X* had not acted. *SH= 1* does not interact or interfere with *BH =1* in this way -- *SH= 1* does not, for example, deflect Billy's rock away from impact. Of course, if *SH=1*, this restricts the possible values of *BH*--assuming (as is a presupposition of the example) that Suzy's rock arrives first, it is not possible for both *SH*=1 and *BH*=1. Once the bottle is hit by Suzy, Billy's rock cannot hit an intact bottle. But it isn't clear that this impossibility is a matter of there being a *causal* relation between *SH* and *BH*.[6]

[5] We emphasize that this is just one of several models of this example that Halpern considers. For reasons of space and relevance we do not discuss his alternative models. The point of our discussion is not to criticize Halpern but rather to draw attention to a limitation in one particular model -- a limitation which will matter subsequently.

[6] After submitting this paper Sander Beckers drew our attention to a rather similar criticism of this feature of Halpern's model in Beckers and Vennekens (2018).

We can provide support for this intuitive assessment by appealing to the following constraint on causal modeling (see, e.g. Woodward, 2020). Prior to writing down the structural equations governing a system, the variables characterizing the system should satisfy the following co-possibility constraint: although the same variable cannot have different values at the same time, all combinations of values for distinct variables (or for the same variable applied to different units) should be possible. Thus, while the same ball cannot have mass of both 1 kg and 2kg at the same time, distinct balls can have the same or different masses and each ball can have a range of possible velocities, given its mass. (Mass and velocity are distinct variables). In physics, such co-possibility constraints are reflected in the state space specified for the system, with the dynamics for the system specified separately and characterizing the causal relations for the system, analogously to the structural equations in causal modeling. We generally think of the possibilities and impossibilities associated with this state space specification as non-causal in character. For example, cannon ball 1's having mass 1 kg does not "cause" it to not to have mass 2kg or "prevent" it from having that mass. Moreover, when variables are such that not all of their values are co-possible, we often take this to be an indication that the variables are not really "distinct" in a way that allows them to stand in causal relationships. To take a well-worn example, if *L* is a variable the values of which correspond to saying hello loudly (=1) or not saying hello at all (*L=0*) and *H* a variable the values of which correspond to saying hello (=1) and not saying hello at all (= 0), then combinations of values like *L=1, H=O* are impossible. This is reflected in our judgment that although these variables stand in a dependency relation of some kind this relationship is not causal[7].

It is arguable that these sorts of co-possibility constraints are violated in the model described above. Assuming an interpretation of *SH* and *BH* as their respective rocks hitting an *intact* bottle, and assuming the usual understanding of bottle and rock behavior in terms of irreversibility, there seems to be no such possibility as *BH=1* and *SH=1*, unless both rocks hit the bottle at the same time, which would turn the case into one involving overdetermination rather than pre-emption. Arguably it is because the co-possibility constraint is violated in the pre-emption scenario when *BH, SH*=1, that it is inappropriate to describe *SH* as causing *BH.*[8] We can bring this out more clearly by contrasting this case with a (science-fictionish) reversible analog: first, Suzy throws, *SH=1*, *BS=1* but then (since Suzy is no longer throwing) *ST=0* and, miraculously, (in accord with reversibility) *SH* and *BS* instantly revert to *0*, so that when Billy's rock arrives an instant later, it strikes a reassembled intact bottle which then shatters (*BH =1,*

[7] Co-possibility constraints of this sort are defended in Halpern and Hitchcock (2010) . However, Halpern has informed us (in correspondence) that he no longer regards these constraints as defensible. Again we lack space for discussion but think that the constraints are reasonable for *causal* relationships.

[8] Another way bringing out the problematic character of the model is that *BH*=0 is ambiguous--it includes both (i) the case in which the bottle has been shattered (by Suzy) and (ii) the case in which the bottle is intact but Billy fails to hit it. If (i), *SH*=0 is impossible, so we have a violation of co-possibility (*BH=0, SH=0* are not co-possible). If (ii), *SH*= 0 and so we can't have *SH=1* causing *BH=0*.

BS=1). This allows for the possibility of both *BH* and *SH =1* in a single scenario in which both Suzy and Billy throw at the same bottle at nearly the same time but of course now the causal structure is very different from what was envisioned in the original scenario. Moreover, it is not so clear that this is naturally viewed as a case of pre-emption rather than two rock impacts having two different, distinct effects. As the contrast between these scenarios makes clear, what distinguishes the two scenarios is whether the causation involved is irreversible.

Note also that once it is recognized that we are dealing with irreversible causation in the original example, a very simple treatment becomes possible. In particular, given the irreversibility of the bottle shattering it follows immediately from *SH=1* and *BS=1* at time *t* that there can be no other cause of *BS=1* occurring after time t. Thus, since Billy's rock hits after Suzy's we can infer that *SH=1* alone caused *BS=1*, without assuming that *SH=1* causes *BH=0*.

It will be instructive to compare this example with another which does involve reversible causation. Suppose that if either of two switches, S_1 and S_2 is on (=1), a light is on (*L*=1). If neither switch is on (S_1, S_2*= 0*) the light is off (*L*=0). Here the causation involved is reversible: the light can be repeatedly turned on and off by changing the switches in the appropriate way. Note that all values of the variables S_1, S_2 and *L* are co-possible, in contrast to the previous example. Suppose *L*=0 before time t, S_1 is turned on at *t*, the light consequently goes on at *t* and then S_2*=1* at a somewhat later time *t+d*. It does not seem intuitive to describe this as a pre-emption case. Prior to *t+d,* S_2 is not in a state that would have caused *L=1* except for the action of S_1 . After *t+d*, it seems most natural to describe the case as involving overdetermination, with both S_1=1 and S_2 =1 causing *L=1*. Note that in this case we do not need anything analogous to the *SH* and *BH* variables in 3.4- 3.6 or the causal relationships represented by these equations. Moreover, although in the dual light switch case overdetermination is possible even if the two causes initially come on at different times, nothing analogous is possible in the Billy/Suzy preemption case—in that example overdetermination is only possible if the two causes *BH=1, SH=1* become operative simultaneously. These differences between the two cases are closely related to the fact that in the light switch case "sustaining causation" must be present if the light is to remain on—at least one switch must remain in the on position. In the bottle case, no such sustaining causation is present—Suzy's rock doesn't have to (can't) do anything after the bottle is broken because (as far as the example goes) there is no possibility of the bottle reverting to its unbroken state.

Now contrast the version of the light case above with the following version which does involve irreversible causation: when either of two switches is on, the light goes on and never goes off regardless of the subsequent position of the switches. S_1 goes on at time *t* and the light goes on at *t* while S_2 is still off—here it is unambiguous that S_1 is the actual cause of the light being on. If S_2 is subsequently on at time *t +d*, it is not an actual cause of the light being on, either at time *t* or later—there is no overdetermination when both are on, as would be the case when there is reversible causation.

Again, the moral is that we need to pay attention to whether the causal relations in scenario are reversible or not. This affects not just how we ought to model but which kinds of pre-emption and overdetermination are possible.

4. Conclusion: Reversibility and Control

In addition to their importance for causal inference and modeling, reversible and irreversible causation also carry different possibilities for manipulation and control. Structures involving reversible causation have the advantage of allowing effects to be turned on and off via manipulation of their causes. This is a desirable feature in many biological systems and is present by design in many artifacts. A regulatory gene that controls protein expression can turn this on and off, thus creating a protein product as needed but also preventing excessive buildup, which is likely harmful. Structures involving reversible causation are also structures that can be reused by returning the cause to its original state. By contrast, irreversible causal relations are associated with cases in which repeated on/off modulation of the effect via the original cause is not possible and in which, no other means for accomplishing this are practically available. In fact, many cases of irreversible causation are associated with damage or destruction, as seen in many cases that figure prominently in philosophical discussion. When a cause acts irreversibly a system is put in a state that it cannot readily get out of and this means that the system is not available for other sorts of uses, as exemplified by a broken artifact.

This leads us to be especially attentive to causes that are irreversible--seatbelt laws, helmet laws, and smoke detector regulations are designed as added protection to prevent the triggering of irreversible causes that lead to death and loss. On the other hand, putting a system in such a state can sometimes be desirable and, if this is so, it may be a good strategy to exploit an irreversible causal relation. When the murderous tyrant is poisoned, it is good that he stays dead and nothing more needs to be done to keep him in that state.

References

Blanchard, T., Vasilyeva, N., and Lombrozo, T. (2018). Stability, breadth and guidance. *Philosophical Studies*, 175(9):2263–2283.

Beckers, S, and Vennekens, J. (2018) "A Principled Approach to Defining Actual Causation" *Synthese* 195:835–862

Halpern, J. and Hitchcock, C. (2010) "Actual Causation and the Art of Modeling" in Dechter, R., Geffner, H. and Halpern, J. (eds.) *Causality, Probability and Heuristics: A Tribute to Judea Pearl*. London: College Publications, 383-406.

Halpern, J. (2016). *Actual causality.* MIT Press. Cambridge, MA.

Kearney, M. and Wilson, R. (2018). Male earnings, marriageable men, and nonmarital fertility: Evidence from the fracking boom. *The Review of Economics and Statistics*, 100 (4): 678-690

Liberson, S. (1985). *Making it count*. University of California Press. Los Angeles, CA.

Ross, L. N. (2018). Causal selection and the pathway concept. *Philosophy of Science*, 85:551–572.

Wilson, W. (1987) *The Truly Disadvantaged*. Chicago: University of Chicago Press.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, 25(3):287–318.

Woodward, J. (2020). Causal Complexity, Conditional Independence and Downward Causation. *Philosophy of Science* 87: 857-67