Models of Causality and Causal Inference

Barbara Befani

Introduction

The notion of causality has given rise to disputes among philosophers which still continue today. At the same time, attributing causation is an everyday activity of the utmost importance for humans and other species, that most of us carry out successfully outside the corridors of academic departments. How do we do that? And what are the philosophers arguing about? This chapter will attempt to provide some answers, by reviewing some of the notions of causality in the philosophy of science and "embedding" them into everyday activity. It will also attempt to connect these with impact evaluation practices, without embracing one causation approach in particular, but stressing strengths and weaknesses of each and outlining how they relate to one another. It will be stressed how both everyday life, social science and in particular impact evaluation have something to learn from all these approaches, each illuminating on single, separate, specific aspects of the relationship between cause and effect.

The paper is divided in three parts: the first addresses notions of causality that focus on the simultaneous presence of a single cause and the effect; alternative causes are rejected depending on whether they are observed together with effect. The basic causal unit is the single cause, and alternatives are rejected in the form of single causes. This model includes multiple causality in the form of single independent contributions to the effect. In the second part, notions of causality are addressed that focus on the simultaneous presence of multiple causes that are linked to the effect as a "block" or whole: the block can be either necessary or sufficient (or neither) for the effect, and single causes within the block can be necessary for a block to be sufficient (INUS causes). The third group discusses models of causality where simultaneous presence is not enough: in order to be defined as such, causes need to be shown to actively manipulate / generate the effect, and focus on how the effect is produced, how the change comes about. The basic unit here – rather than a single cause or a package – is the causal chain: fine-grained information is required on the process leading from an initial condition to the final effect.

The second type of causality is something in-between the first and third: it is used when there is no fine-grained knowledge on how the effect is manipulated by the cause, yet the presence or absence of a number of conditions can be still spotted along the causal process, which is thus more detailed than the bare "beginning-end" linear representation characteristic of the successionist model.

1. Simultaneous presence of cause and effect: the successionist view

This paragraph covers some of the more traditional accounts of causality – based on both regularity / invariance of patterns and counterfactual thinking. Because the basic causal unit considered here is the single cause, most quantitative, variable-oriented methods are based on this model; including regression[1], experiments (RCTs) and quasi-experiments..

1.1 Regularity

Historically, the first modern account of causality revolved around the observation of regularities: if potential cause C and effect E are always found together, then either C causes E, or E causes C. The assumption is that a true cause does not work by accident, but operates constantly and regularly, producing the same effect over time and in different settings (hence its characterization as "lawlike").

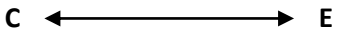
[1] Regression and analyses of correlation are based on Mill's Method of Concomitant Variation, which can be regarded as an extension of the two methods that will explored in detail here: the Method of Agreement and the Method of Difference.

While labeling this a "successionist" view of causation, Ray Pawson (2007) provides us with a visual representation: the "omnipresent arrow" inevitably leading to the effect.

What matters in regularity is the simultaneous observation of two separate entities, while the description of the causal connection (the nature of the "arrow"), or the process leading from C to E remains unknown; what happens in-between cause and effect, what the cause does in order to produce the effect, is kept closed inside what much literature has called the "black box". In Hume's words: "we can never penetrate so far into the essence and construction of bodies as to perceive the principle, on which their mutual influence depends" (Hume, Treatise II, III, I).

In the successionist view, the cause is both necessary and sufficient for the effect: sufficient because all events where the cause is observed also present the effect; "we may define a cause to be an object, followed by another [the effect], and where all objects similar to the first are followed by objects similar to the second" (Hume, Enquiry, VII, II). But also necessary in that "if the [cause] had not been, the [effect] never had existed" (Hume, Enquiry, VII, II).

We may thus redesign the arrow to account for the biunique character of the relation cause-effect:



In mathematical terms, we could say there is an isomorphism from the set of causes to the set of effects.

1.1.1 How causation is claimed: by agreement

Within the regularity framework, causation is claimed through observation of regular co-presence of both cause and effect. In order to claim necessity, the cause must always be present whenever the effect is. And in order to infer sufficiency, the effect must always be present whenever the cause is.

In his "Method of Agreement" (A System of Logic, 1843), John Stuart Mill provides a procedure to establish necessary and sufficient causal links by systematically comparing events on a number of characteristics. If the cause is found to be co-present / connected with two different effects (in two different events), the cause is rejected on the grounds that it is not sufficient (sometimes it produces one and sometimes the other). Similarly, if two different causes are observed together with the same effect (in two different events), then both causes are rejected on the grounds that neither is necessary for the effect. However, if a single cause is always observed together with the same effect, and all other elements can be rejected as potential causes because they are observed in only some events but not all, then it can be inferred that the relation is causal. Below two events are compared and implications are drawn.

C E^1 | C E^2 => cause is rejected because it's connected to two different effects (not sufficient)

C^1 E | C^2 E => cause is rejected because it's connected to effect only sometimes (not necessary)

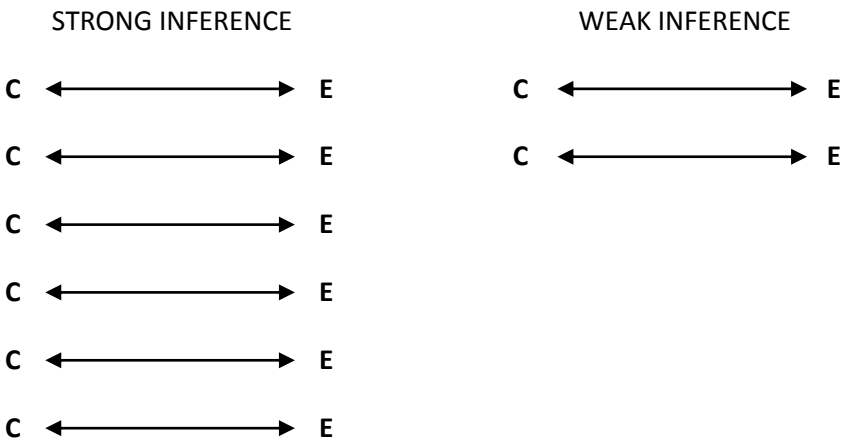
C E | C E => causal link is established but =>

- ⇨ f g h C E | i j k C E all other elements must be present in only one of the events compared, otherwise they could be part of cause or effect

The main problem with the method of agreement lies in checking for the difference of "all possible elements": in practice this requires comparing a high number of events taking place in a wide range of settings. Statistical causal modeling can be suitable for this, being based on the assumption that, after regularity between cause and effect is found in a certain number of cases, the residual characteristics can be attributed to "chance" and thus must necessarily vary in a way that cannot be causally connected to the effect.

Because we cannot reach absolute certainty of inference, we have developed ways to evaluate the "quality" (likeliness) of the causal link: as in statistical modeling, within the more general regularity framework, the frequency of association between cause and effect strengthens the causal assumption; and theoretical relationships / associations / models that apply to a high number of cases are "more reliable" in their "explanatory power" than theories applying to a small number of cases or only one case.

In other words, **the strength of the causal association** increases with the number of cases where conjunction between cause and effect is observed; and finding cases where the cause is not simultaneously observed with the effect weakens the inference. In regularity causation, the closer we are to a "law" (see also Hempel's deductive-nomological model, Mill's Method of Concomitant Variations and Davidson's "dose-response" proportionality), the better.



In **impact evaluation**, the regularity approach is useful when the knowledge gap one wants to fill concerns the number of beneficiaries that present given characteristics, say, after an intervention. It does not provide answers on why the beneficiaries have these characteristics, nor on how they were developed following the intervention. Regularity does not allow the evaluator to trace the process leading from cause to effect, and the attribution of the impact, while shown to hold in many cases, lacks "depth" on how the causal association happens.

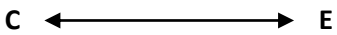
1.2 Counterfactuals

Although regularity and statistical significance are fundamental pillars of scientific enquiry, they only "vouch" for the causal claim and are never able to "clinch" it (Cartwright) with certainty: it's indeed impossible to claim to having considered all possible existing elements of reality in the comparison between events. In order to solve this problem, a method (and more generally, a way of reasoning) has been proposed in which the comparison is not done between highly different cases only sharing the cause and the effect but between identical cases differing only in cause and effect.

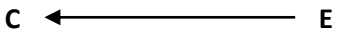
The roots of counterfactual thinking are to be found in Mill's Method of Difference (a.k.a. a double application of the Method of Agreement, see Ragin). Weber (1906) is being influenced by this model when he argues that unlike historical science, social science ought to be answering questions like "what would the course of history have been like if Bismarck had not decided to go to war in 1866": by comparing a real, factual situation (Bismarck decided to go to war) with a counterfactual one (Bismarck decided not to go to war) one should be able to imagine the difference between the consequences of Bismarck's real decision (war) with the counterfactual consequences of Bismarck's counterfactual decision (no war), in order to estimate the "impact" of Bismarck's decision on the course of history.

Counterfactual analyses share several properties with studies aiming to prove regularity:

- both focus on the simultaneous presence of a single cause with an effect, without enquiring into the nature of causal relation (the process, or the "arrow");
- both see the cause as both necessary and sufficient to produce the outcome



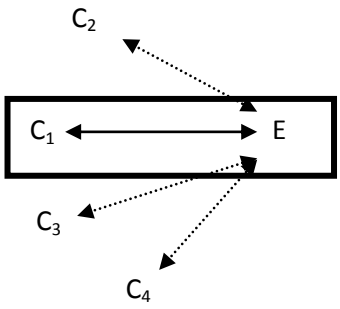
The property of sufficiency has been dropped by Lewis (1973), who argued that in order for a causal link to subsist, the following had to hold: a) when C occurs, then E occurs; b) when E occurs, then C occurs; and c) when C does not occur, then E does not occur. Lewis argued that the fourth proposition "when E does not occur, C does not occur" did not need to hold: in other words, the cause is not sufficient, because it can occur also when the effect does not occur; however it is still necessary, because whenever the effect occurs so also does the cause (but, as we will see below, this stratagem is still unable to properly address the multiplicity of causation).



1.2.1 How causation is claimed: by difference

Apparently, counterfactuals have the advantage of needing only two events (as opposed to infinite) to infer causation; however, this advantage is only apparent because those two events need to be identical on an infinite number of elements except cause and effect.

The cause is isolated through the careful choice of the two events to be compared:



In Mill's Method of Difference, causality of C with regard to effect E is claimed in the following way:

f g h i j k | f g h i j k C E => C is the cause of E (or E the cause of C)

The above two events are compared and C is the only "new entry" in the second event: all the other elements f g h i j k l m are present in both.

When other factors are present in only one event along with C and E, we cannot infer causation; for example in:

f g h i j k | f g h i j k L C E => either L or C could be causes of E.

While f, g, h, i, j and k are rejected on the grounds that they are present in both events, L cannot yet be rejected. In order to reject L, too, we need to find the case f g h i j k L.

As with the regularity approaches, the **strength of inference** through counterfactuals increases as the number of alternative causes we are able to reject increases; the higher the number of elements that can be shown to be equal in the two events, the better.

f g h | f g h C E => **WEAK INFERENCE** (i, j, k, l, m and n haven't been rejected yet)

f g h i j k l m n | f g h i j k l m n C E => **STRONG INFERENCE** (many more causes have been eliminated)

A famous strategy used to find a specific event presenting a number of specific factors without C and E is the experiment in controlled settings (see paragraph 3.1). In **impact evaluation**, a number of strategies are used to design experiments and quasi-experiments: RCTs, statistical matching, regression discontinuity, difference-in-difference, etc. These are usually considered the strongest, most rigorous and most robust methods to attribute a given result to an intervention. However, even when the known threats to experiments are controlled for (see paragraph 3.1) and the causal association covers a high number of cases, a knowledge gap remains on the characteristics of the association between a given factor and a given effect. Like the regularity approach discussed above, counterfactuals do not provide answers as to what happened "between" the alleged cause and the effect; eg. on what the "true" cause was at the micro, in-depth level.

1.3 Critiques to the successionist view

The main critiques that have been addressed to the successionist view concern direction of causation, the nature of the causal relation, and the interaction with other causes. Indeed, neither the regularity and

counterfactual approaches – although enjoying different comparative advantages in rejecting alternative explanations – enlighten on these aspects specifically.

1.3.1 Direction of causation

With the partial exception of Lewis's stratagem and a number of sophisticated econometric techniques checking for the goodness of fit of various models, none of the approaches addressed above provides convincing ideas, methods or criteria to deal with the direction of causality. In most problems, C and E could be either causes or effects: there is a biunique relation (isomorphism) between the set of causes and the set of effects: only one effect for each cause, and only one cause for each effect. In simple (eg univariate) regression, the dependency relation is purely contingent: $y = ax + b$ could instantly become $x = y/a - b/a$. In multivariate modeling, the functional form can be changed so that y becomes one of the independent variable and one of the x-es becomes dependent[2].

Similarly, the mere comparison of two identical events differing only in one feature, although it can draw attention to the simultaneous presence of the two different states and lead observers to investigate whether one transformed (into) the other, does not provide any insight on the details nor the direction of this transformation.

In order to remedy, Hume (1739) has proposed the criterion of temporal precedence: when confronted with two events that are one the cause and the other the effect, the event preceding the other is to be identified with the cause, and the event preceded by the other with the effect:

$$\mathbf{C, t_0} \longrightarrow \mathbf{E, t_1}$$

However, temporal precedence has its limitations. First, it is not always easy to locate events temporally with a high degree of precision: Weber's argument that the Protestant Reform caused the development of capitalism requires that the Reform precedes Capitalism; however, it is not easy to say with precision when the Protestant Reform became fully institutionalized, nor when did Capitalism, and thus whether the former preceded the latter (Brady 2002). Secondly, even when events are located correctly and precisely along a time scale, the actual cause might take place in a different point of the time scale than when those events happened. For example: when women leave full-time schooling in order to get married, we don't know whether marriage is the cause or the effect of their decision to leave school, even though school leaving always precedes marriage. They could have decided to get married either before leaving school, or later, even though marriage comes always last on the time scale (Brady 2002, Marini and Singer 1988). In this case the actual cause is not "school leaving" but "decision to get married". In symbols:

$$\mathbf{School\ leaving, t_0} \longrightarrow \mathbf{Marriage, t_1}$$

$$\mathbf{Decision\ to\ get\ married, t_0} \longrightarrow \mathbf{School\ leaving, t_1} \longrightarrow \mathbf{Marriage, t_2}$$

$$\mathbf{School\ leaving, t_0} \longrightarrow \mathbf{Decision\ to\ get\ married, t_1} \longrightarrow \mathbf{Marriage, t_2}$$

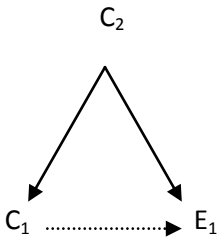
[2] A number of techniques can be used to assess the goodness of fit of the different models, providing information on direction; however this "directionality" is limited to quantitative relationships among variables and is based on the regularity with which a number of numerical contributions add up to a total effect.

In **impact evaluation**, sometimes interventions are part of a causal flow that can present different shapes, including recursive: as in when they are made possible / implemented thanks to the presence of skills or demands that the intervention itself is supposed to generate or improve. Just because the intervention happened before a result, it doesn't mean it has *caused* it: the "real" cause can be something that produced both the result *and* the intervention. Which brings us to the next paragraph.

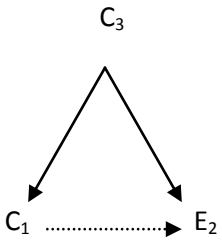
1.3.2 Correlation or causation?

Every social scientist knows that correlation is not the same as causation: and that when we say that C is correlated with E, there might be a third factor which actually causes both C and E. This problem – also known as spurious correlation – is approached in statistical modeling through the assumption of independence of variables and the isolation of the "pure", independent contribution of each single factor in the model.

This category of models solves the problem of correlation and opens that of independence, which we will deal with below. Here we will address the problem of correlation by remarking that, while counterfactuals apparently do not have this problem because all variables in the model are "held constant", in fact the action of the cause could be dependent on a particular context that both the treatment and the control cases are embedded in. Therefore, even when both terms of the comparison are perfectly equal, in fact their behavior might be linked to some contextual feature C_2 that influences the reaction of both; and affects C_1 on one hand and E_1 on the other.



What the counterfactual can thus illuminate on is the effect in a particular situation (E_1); which cannot be generalized to a context C_3 where the behavior / effect of the same cause might be different (E_2).



Cartwright notes that counterfactuals are able to tell us whether something works "here", in a particular context, at a specific time and place; but fall short when we ask whether the same works "somewhere else", at another time and place; let alone everywhere and all the time. In **impact evaluation**, even when experimental techniques like randomization ensure equivalence of the two terms of comparisons, the contextual characteristics in which both control and treatment groups are embedded in might influence their response to the intervention, thus impairing the external validity of the experiment (see paragraph 3.1).

In sum, the causal relation inferred through both regressions and counterfactuals can be spurious.

1.3.1 Can causes be independent / unconditional?

Even in situations where the above-discussed threats are controlled, the successionist view is unable to conceive of causes as behaving differently in different settings, because it postulates causal independence and the interaction of causes is usually not addressed. Causes are modeled as independent forces whose behavior is not influenced by context, experimental settings, or other causes. In other words, they are mostly **single** causes acting **independently / unconditionally**. Their power / influence is constant, does not depend on other factors nor varies according to them, and is defined in terms of coefficients of concomitant variation: that is, their causal power defines how much the value of the effect will change, on average, due to the presence of one more unit of the cause.

The coefficient is the closest the successionist approach has to a description of the causal process (the "arrow") and – being an average – is constant throughout the replications of the phenomenon: the relation between cause and effect (embodied by the coefficient) is thus not allowed to change according to the context. The context, if considered, is modeled in turn as an independent force with its own "adding-up" power to the "total". The process is conceived as being linear.

Multiple causation in the successionist view

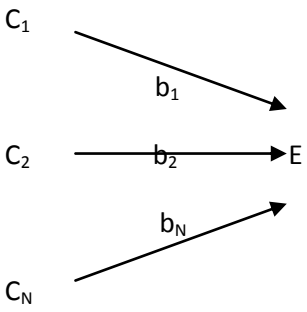
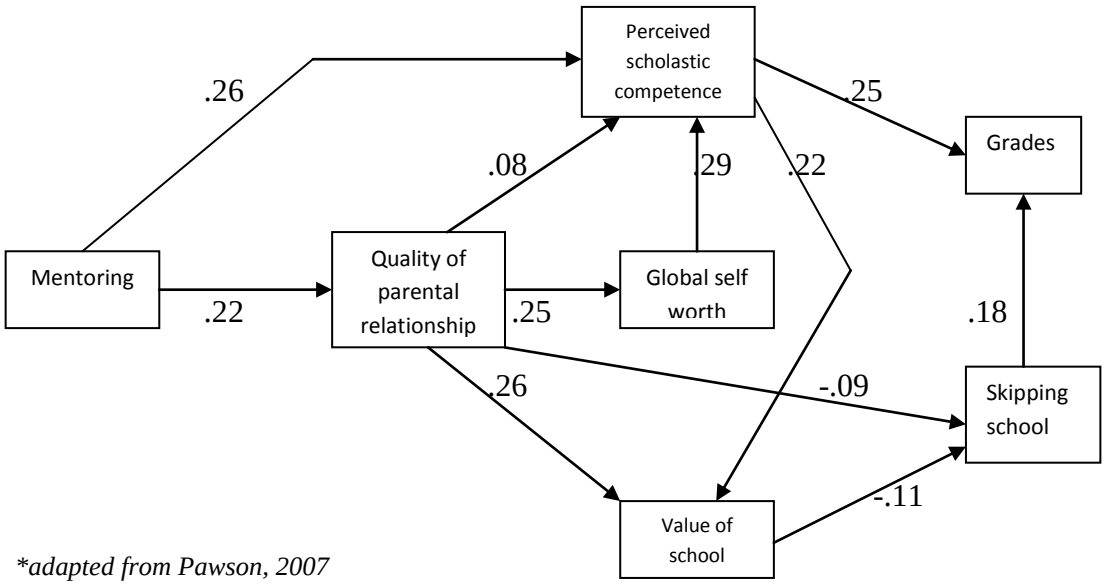


Figure 1: Path model of direct and indirect effects of mentoring*



adapted from Pawson, 2007

But context can in fact change the power of other causes in ways that it does not necessarily make sense to add up and make an average of: as the figures above show, C_1 ends up not being sufficient for neither E_1 nor E_2. Does this mean C_1 is a cause of neither E_1 nor E_2? The successionist view does not help in deciding whether we should still call C_1 a cause of E_1 or E_2 because it is suited to provide information on average quantities and is not tailored to perform an analysis of necessity or sufficiency of single causes or packages or multiple causes; as are the configurational methods expounded in the next paragraph.

2. Co-presence of multiple causes: necessity and sufficiency

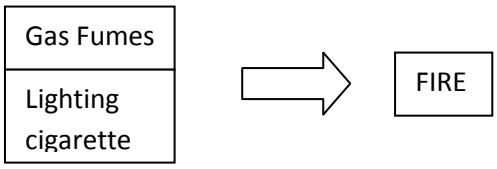
The above-mentioned Methods of Agreement and Difference can be considered logically equivalent but have different comparative advantages: when two events seem highly different, one will tend to spot their similarities; conversely, when they are highly similar, one tends to spot their differences. But this choice varies across individuals: when confronted with the same event, an individual might think of a highly similar event and spot the difference, while another might prefer to compare it with a highly-different one and spot the similarities. In a maternal health intervention in Tanzania, an expert in maternal health will tend to observe the differences between this and many other similar interventions she has knowledge of; while someone with, say, country-expertise in Tanzania will tend to spot a small number of similarities between this and the many different interventions from all sectors he has knowledge of.

But even when we compare events we want to explain with an identical event where the cause did not take place, say a similar area in Tanzania that did not receive assistance for maternal health, this comparison (or control) event might be *equal in different ways* to the treatment one. The similarities identified between the two events might differ across individuals and situations. In other words, these similarities are "chosen" according to the available evidence and knowledge; and different evidence / knowledge might be available to different individuals.

Randomistas and counterfactualists will immediately think that this is an argument to suggest controlling for a higher number of factors; but this is not the point. The point is not that the highest number possible of characteristics should be compared, but rather that in everyday life we are not interested in the average effect of causes when we attribute causality: we mostly just easily figure out what caused what, when, and depending on the situation we attribute different causes to similar effects.

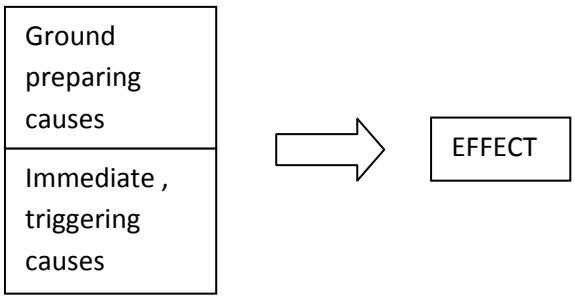
For example, consider the situation of a person lighting a cigarette at home in the presence of gas fumes accidentally leaked from the heating system and thus causing a fire. Because s/he has often lit a cigarette at home without causing a fire, s/he will tend to think of the control event "lighting a cigarette at home in the absence of gas fumes" and blame the presence of gas fumes for the accident. But if the same accident happens at a gas station, the same person will tend to think of all the other times she was at a gas station, with gas fumes around, and didn't light a cigarette: with this control event in mind, she will blame the act of lighting a cigarette for the accident.

Now, what is the real cause of the event "fire"? The immediate cause is lighting the cigarette, but it's obvious that the fire would not have happened without the gas fumes: the same person in fact blames both factors albeit separately and in different situations. The answer is that both are causes of fire, because fire would not have happened without either: so, although no cause taken singularly is sufficient, both are singularly necessary and jointly sufficient for fire to happen.



One could observe a difference between the causes, in that gas fumes are a sort of "background", ground-preparing cause, while the cigarette lighting is the immediate cause of fire. This distinction is even more evident when inferring causation for historical events. Suppose two historians need to evaluate the causal significance / impact of the assassination of the Archduke of Austria in relation to the onset of World War I. One is an expert in political instability, the other an expert in assassinations. The latter will immediately think of a number of assassinations that had negative consequences and claim that the assassination of the Archduke did indeed cause WWI. The former will think of comparable situations of political instability, all of which eventually led to a war, with or without an assassination, and claim that the assassination was not really a cause of WWI, but political instability was (Brady 2002).

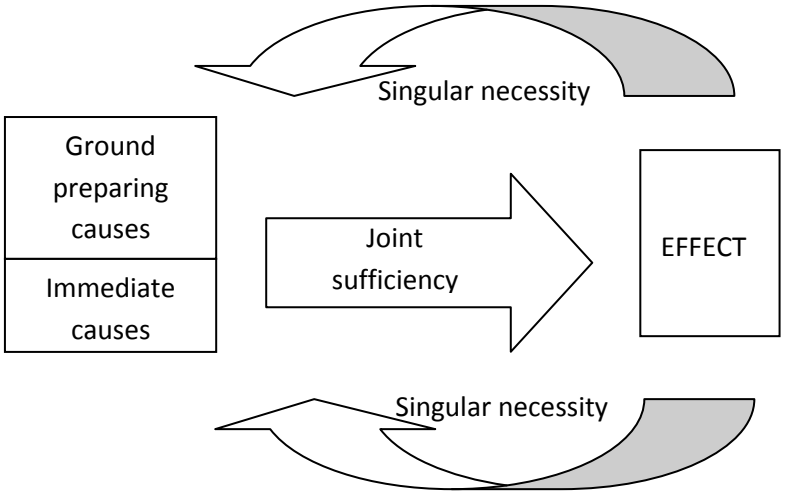
Both are right: they just focus on different types of insufficient causes. The assassination expert is focusing on the immediate cause of WWI, while the political instability expert is focusing on the "ground-preparing" causes. The assassination is akin to cigarette lighting; and political instability to gas fumes.



Historical processes, including development, can be described as the accumulation of a number of ground preparing causes, "lit" now and then by "sparks" that cause leaps and bounds, eg. in the development level of countries and regions. Ground preparing causes alone are not sufficient: they need a "trigger" without which they can't produce the effect; but the trigger alone is not sufficient either, and would not "perform" without the other causes.

Interventions can sometimes prepare the ground, or other times provide the trigger and achieve effects immediately. For example, immunization immediately reduces infection rates; and direct food assistance immediately reduces the number of starvation deaths. However, not all causes act so quickly: in order to acquire a good education a child must have books, good teachers, some incentives to invest time and energy in studying, and ensure that she attends school regularly. In some cases results might be poor because only 2 or 3 causes are present; but when the third or maybe fourth cause is present results might start improving – not necessarily in a linear way, but possibly in a "from no to yes" kind of way. From poor performance to excellent performance. **Impact evaluations** that attribute the effect entirely to the last cause on the grounds that it was the only one always or inevitably present in all cases risk overestimating its power, while at the same time underestimating the importance of the other, ground-preparing causes, without a combination of which the trigger would have not performed.

The ground preparing causes are thus not sufficient but are nonetheless necessary: gas fumes flying in the air do not catch fire without a spark; but when the fire happens, the presence of gas fumes (or of other combustible agents) is always detected. The immediate causes are also not sufficient but necessary: lighting a cigarette in a gas fume-free environment does not cause a fire (unless some other underlying condition is present), but when the fire happens, so is the cigarette lighting detected.



This notion of causality is relevant to **impact evaluation** because although intermediate outcomes might not be sufficient to produce a specific effect:

- They might be proven to be necessary
- The combination of intermediate effects with other supporting, ground preparing and / or sustainability factors can be shown to be *jointly sufficient* for impact.

Back to the schooling example: if evaluated with the regularity approach, the situation described by the table below would conclude that all causes on average have a similar effect on performance, ignoring the fact that – far from having an independent influence – none of them is able to have an impact without ALL the others being present.

When taken one by one with a counterfactual approach, each cause would appear miraculous, because everything else staying the same, each cause would be capable of increasing performance from low to high. But in a situation where an intervention would provide, say, only books, then only the two blue-colored cases would be compared: what would matter in this case is that "everything else stays the same", not what the other causes are. The results would be entirely attributed to books, ignoring the fundamental contribution of teachers, regular attendance and incentives.

Books	Good teachers	Incentives	Attendance	Performance
YES	YES	YES	NO	LOW
YES	YES	YES	YES	HIGH
YES	NO	YES	YES	LOW
YES	YES	YES	YES	HIGH
NO	YES	YES	YES	LOW
YES	YES	YES	YES	HIGH
YES	YES	NO	YES	LOW

YES	YES	YES	YES	HIGH

When looking at the bigger picture, one soon realizes that "everything else staying the same" is not enough informative in this case, and once the interaction of causes is disentangled and understood, it becomes clear that without many other "ground preparing causes" what the intervention provided (books) would not be able to reach the critical mass / diversity of resources needed to "climb the step" and finally impact performance. The above case is much better understandable with a non-linear, discrete / step-like, necessity-sufficiency analysis that with a successionist, linear / continuous approach to causality looking for attribution to a single cause or to estimate the independent effect of single causes.

At the theoretical level, non-linear multiple causality is exemplified by a famous type of cause introduced by John Mackie (1974).

2.1 The INUS cause and the "causal field"

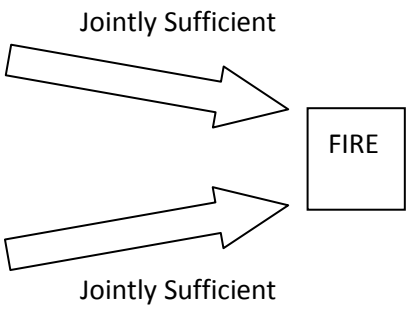
In an attempt to remedy the inadequacy of the successionist view in accounting for the effects arising from interaction of causal agents, John Mackie introduced in 1965 the notion of "causal field", exemplifying the idea that the link to be studied is not between the effect and a single cause, but between the effect and a causal package: a "block" of single causes that might not have an independent influence on the effect. In 1974 the same author theorizes a special type of cause called the INUS: an insufficient (I) but necessary (N) part of a causal package, which is in itself unnecessary (U) but sufficient (S).

Fire cannot only be caused by gas fumes and cigarette lighting, although they are jointly sufficient for it. It can also be caused by, for example, striking a match on a red phosphorous surface. Each of these four causes is an INUS: none of them is sufficient for fire; but each of them is necessary for a combination to be sufficient for fire. The match in itself does not light a fire, but neither so does the red surface: none of them alone are sufficient, and both of them need the other to produce fire. In other words they are jointly sufficient in that they are part of a sufficient combination. This combination, however, is not necessary for fire to happen: fire also happens when cigarette lighting devices being activated meet gas fumes.

THE INUS CAUSE

Gas Fumes	Necessary	Insufficient	Jointly Unnecessary
Lighting cigarette	Necessary	Insufficient	

Striking a match	Necessary	Insufficient	Jointly Unnecessary
On a red phosphorous surface	Necessary	Insufficient	



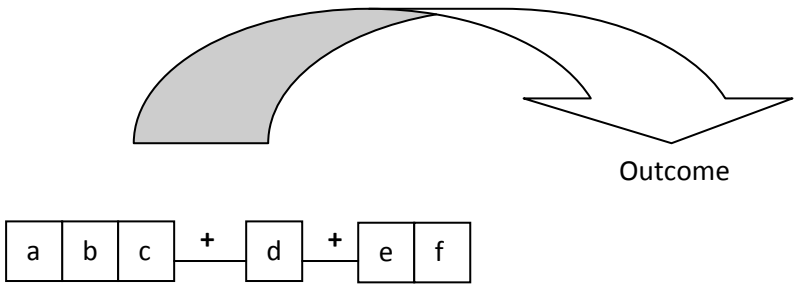
The INUS cause has been associated with contribution analysis (NONIE 2), and the idea that an intervention strategy might not be necessary to achieve a result which can be obtained in multiple ways; however, an intervention can be shown to be a necessary component of a sufficient strategy, and thus be shown to "cause" it, in combination with other factors. When considering the different paths that produce the result, it can also be argued that the one including the intervention – although not the only one / not necessary – is preferable (eg because it speeds up achievement or has lower costs in terms of social justice) or the only possible one in a specific context (eg with low economic performance, poor institutional capacity, etc.) (see Rogers in Forss, Marra and Schwartz eds., 2011)

2.2 Configurational causation

The analysis of necessity and sufficiency thus reveals that many causes come in "packages": pairs or vectors of causes, and are unable to produce the effect unless they are combined with other causes; and that the same effect could be produced by several different combinations of different single causes.

In what is called the configurational view of causation (Pawson 2007, Ragin 1987, Rihoux & Ragin 2008), a cause is thus identified with a configuration / constellation of conditions, or more precisely with a combination of single causes producing the same effect. Different combinations may lead to the same outcome; and similar combinations may lead to different outcomes, because individual conditions can affect the outcome in opposite ways, depending on what other factors they are combined with.

Figure 2: Configurational causation



2.2.1 Critiques of configurational causation: component causes of mechanisms

Configurational causation solves some problems connected with the successionist view but not all. Causality is no longer essentially single and independent, and is now seen as properly multiple and conjunctural; however the problem of establishing the direction of causation is not yet satisfactorily addressed. Moreover, the correlation problem is replaced by a different "association" problem: it is no longer a matter of observing the simultaneous presence of cause and effect, but the simultaneous presence of multiple single causes and the effect. The black box has become partially transparent and some characteristics of the causal process begin to emerge (we know more than just beginning and end, we do have some inbetween elements), but the understanding is not yet fine-grained and combinations need to be interpreted. We have started to "peek in" but still haven't opened the black box. We can glimpse some characteristics of the shaft of the causal "arrow" but we likely need more details in order to establish direction and connect specific parts of the cause and specific parts of the effect in a finer way (see pairing and pre-emption below).

From the schooling example we know that in order to achieve good results in school, children must have books, attend regularly, have incentives to invest time and energy in studying, and good teachers. However, a list of causal conditions does not provide information on *how* those conditions are connected, what should come first or later, and whether there are any synergies between factors, how does one help the other, etc. A better description of the causal chain resembles an argument of the kind "books increase the chances of reading written text and thus improve reading ability" or "good teachers speed up the learning curve by adapting their method to the learning possibilities of children", "families provide incentives by to the child by collaborating with teachers", and so on. Mechanisms (see next paragraph) describe at a high level of detail how each cause or package of causes manipulates / changes the situation in order to produce the effect.

3. Manipulation and Generation

While everyone might know the recipe to a meal, just having some ingredients on the table or in the fridge does not make the meal. Someone must actually put together the ingredients in a certain way to obtain the final effect – the meal. While the configurational view of causation sheds some light on necessity and sufficiency (the recipe), this paragraph focuses on a notion of causality that informs on how the ingredients must put together; following what order and techniques. In other words, we will address the specific processes of mixing, mashing, flavouring, cooking, etc. that can make different meals out of the same ingredients depending on how someone mixes them together.

This should provide enough clarity on the problems of asymmetry and direction: a.k.a. it's not the meal that makes the ingredients, but some individual that uses the ingredients to make the meal. It does so in two ways: first by exploring the notion of manipulation and controlled experiment; and later by introducing the concept of mechanism as an actual description of the causal process taking place between cause and effect (the "arrow").

3.1 Human agency: claiming causation through intervention

In the first paragraph, causation was described as the simultaneous presence of cause and effect: the presence of two separate, detached entities at the same time in the same place. A number of authors, however, describe causality as "forced movement: […] the manipulation of objects by force; […] the use of bodily force to change something physically by direct contact in one's immediate environment" (Lakoff and Johnson, 1999). Rather than being detachedly / shyly "just" present together in the same event, causes "bring, throw, propel, lead, drag, pull, push, drive, tear, thrust, or fling the world into new circumstances" (Brady 2002). The emphasis here is on intervention and agency, and "the possibility that the failure to engage in manipulation will prevent the effect from happening" (Brady 2002).

Indeed, counterfactuals are sometimes offered as "explanatory laments" or as "sources of guidance for the future" (Brady 2002), such as "If he had not had that drink, he would not have had that terrible accident". In a situation where a person chooses a new, faster road to get home from work, and gets their car hit by a drunk driver going too fast, the person will tend to think "next time I'll take the old road on the way home", identifying the cause of the accident with her/his choice of itinerary; rather than think "there should be stricter laws on drunk driving", identifying the cause of the accident in the wider legal system. Research has shown that individuals tend to infer causal power to those events that can be manipulated: "when

considering alternative possibilities, people typically consider nearby worlds in which individual agency figures prominently" (Brady 2002).

The idea of manipulation permeates the entire world of public policy interventions: by allocating resources and planning and implementing a number of activities in a certain context we hope to "change" something. And **impact evaluation** aims at understanding whether we have succeeded and provoked, *caused* some effect. The two main ways to evaluate our causal power are a) organizing experiments and b) put our actions under the microscope in order to *see* causality at work – eg. what part / detail / component / cell / molecule of our action changed what part / detail / component / cell / molecule of the affected reality to produce the effect (biologists do both at the same time). We address the former way in this paragraph and the latter way in 3.2.

Manipulation is an important addition to counterfactual thinking in that it can solve the asymmetry problem: in laboratory and controlled-settings experiments, the scientist can manipulate the cause, activating and disactivating it at will, and collect empirical evidence on the consequences. In randomized experiments, the researcher administers the treatment to only one of the two identical groups, and collects evidence on the treatment consequences. In these cases there is no doubt on the direction of the causal arrow: however, manipulation does not protect causal inferences from other risks.

Experiments are indeed subject to three main forms of criticism: lack of external validity, limited applicability (eg threats to internal validity) and pre-emption.

3.1.1 Critique #1: Lack of External Validity (Accidentality)

Causation inferred through experiments can be considered "accidental" because it might be independent of any law ("possible worlds are not considered", Brady 2002) and does not provide evidence on the behavior of the cause outside of experimental settings (Cartwright 2012, Rothman 2005, Howe 2004). Its value as "guidance for the future" is doubtful when, even though one successfully avoids taking the new, faster road on the way home, perhaps more drunk drivers are around because that road happens to be full of bars and pubs, and accidents are still likely to happen. Similarly, although human agency can successfully intervene in determining that no cigarette be lit, (see example above), gas fumes in the house are still dangerous because other sparks can easily fly (gas cookers in the kitchen, heating stoves, etc.).

Generalization based on experiments stand on the (sometimes risky) assumption that the relevant contextual conditions in which the experiment takes place will remain unaltered throughout the reality targeted by generalization (eg basic biologic properties of the human body that do not change through individuals and interact with a given drug always in the same way). As Cartwright puts it, "experiments tell us that an intervention works here" but "we need something else" to extrapolate that finding and deduct that the same intervention will work "somewhere else".

3.1.2 Critique #2: Threats to Internal Validity

Although RCTs are good at removing selection bias through randomization, they are still exposed to post-selection differential change: after selection, change might take place in a differentiated way in the two groups. In particular (Scriven 2008):

- Groups might be differentially influenced by their awareness of being part of the treatment or the control group (a.k.a. the Hawthorne effect);

- Subjects in the treatment group might leave the experiment for reasons related to the treatment, which are not the same reasons that subjects in the control group would leave the group for (a.k.a. differential attrition / mortality);
- The two groups might interact and influence each other in an asymmetrical way: for example subjects in the control group might start imitating subjects in the treatment group (a.k.a. diffusion effect).

3.1.3 Critique #3: Pre-emption

While successfully solving the direction problem, manipulation (or full-on experimentation) does not account for those causes that could have acted, but where prevented from acting by the cause that actually operated. The following example is often presented in the philosophical literature (Brady 2002): a man walks across a desert with two enemies chasing him. The first enemy puts a hole in his water can. The other enemy, not knowing about the first, puts poison in his water. Manipulations have certainly occurred, and the man dies on the trip. Now, the first enemy thinks he caused the man's death by putting the hole in the can. Similarly, the second enemy thinks he himself caused the death, by putting poison in the water. In reality, the water dripping out of the can might have prevented the cause "poison" from acting: and the man might have died of thirst rather than from poisoning. So the counterfactual "if the second enemy had not put poison in the water, the man would have survived" doesn't hold, because the man would have died anyway of thirst.

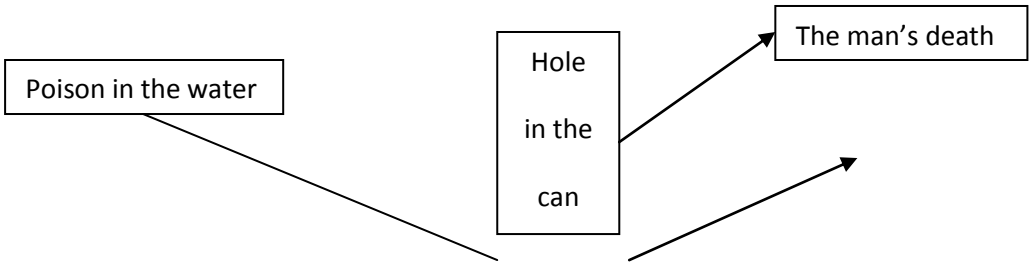
Let us assume the man died of thirst. The closest possible world to the one that has actually happened is one where poison would have killed the man anyway: so a well-constructed experiment on the impact of a hole in the can on the man's health would conclude that the hole did not kill the man, because *ceteris paribus* the "control" man also died. But this is incorrect because the "treatment" man did in fact die from the scarcity of water caused by the leak, so the hole did have an impact, and is certainly able to kill another man in the same situation.

In policy interventions, targeted groups sometimes have different strategies / opportunities and before acting they wait to see whether they have been selected for a certain treatment. If they are not selected, they will act in a way that compensates for non selection, and perhaps might achieve similar results to the selected. In this case an **impact evaluation** that compares treatment and control groups might conclude that the intervention had no net effect or was not "necessary" for the result; but it cannot conclude that it did not have some causal power in producing the effect, because the treatment group achieved the result through the intervention, benefiting from the resources provided by it and not by choosing and alternative strategy like the non-selected / control groups. It's not possible to say that the intervention did not cause / produce the outcome, but it's possible to say that while exerting its causal power, it prevented another potentially effective cause from operating.

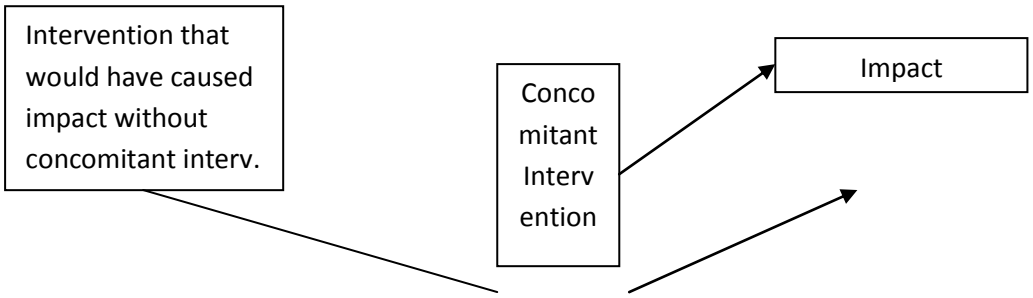
Similarly, when an intervention is implemented in contexts with different previous and current interventions interacting with it, it might seem to work differently, eg only in some contexts, because its operation might have been "displaced" (prevented) by other interventions that acted first. Similar concepts are the "crowding out effect" in economic and social science theory (see for exampler Elster 1998), and path dependence (Page 2006).

Metaphorically, when pre-emption operates the causal arrow is "broken" and displaced by another causal arrow:

Pre-emption (crowding-out, path depedence): the man in the desert



Pre-emption (crowding-out, path dependence): impact evaluation



In the above case it might appear easy / attractive to dismiss the intervention as useless or irrelevant. When administrations are interested in the pure effect (eg whether the man dies or not, or whether an intervention worked in a specific time and place or not, without knowing why), experiments might rank the intervention low or zero on impact score; however, an intervention might not work in a specific setting where other interventions were taking place, but work well in other, less-crowded settings. This does not mean that the intervention doesn't have the causal power to produce the effect; it just means that it needs specific conditions to do it. Discovering how the intervention produces the effect might explain why it works better in some cases than others.

3.2 Generative causation: the description of the causal mechanism

Pre-emption is part of a more general problem that all approaches to causal thinking expounded so far are unable to address: the pairing problem. In an example above, although school leaving is consistently found to precede marriage, it cannot be considered a cause of marriage: it's difficult to argue that people get married because they suddenly find themselves out of school. At the same time, no one gets married without having previously decided to marry: so the decision to get married can be considered the "real cause"; and a closer examination reveals that the decision can be taken either before or after leaving school.

In many cases, even if constant conjunction between cause and effect is proved; even if the cause is articulated into a configuration; even if direction is established and even if the ineluctability of the effect is demonstrated, we still don't have enough information to properly "pair-up" cause and effect. Even if we know the man in the desert would have died anyway, we still can't attribute that particular death to a

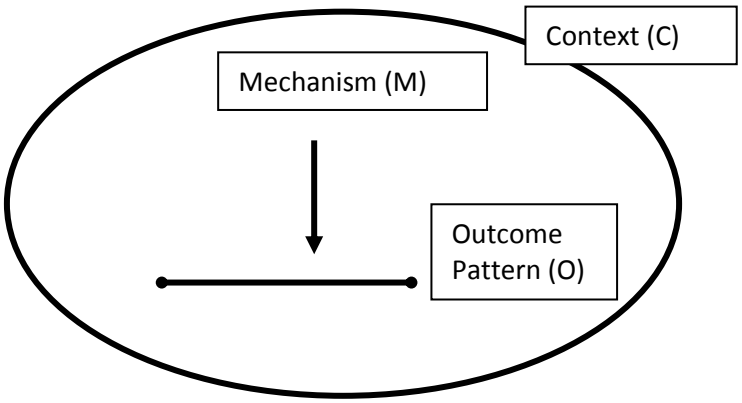
particular cause (leak or poison). We might know that an intervention clearly works, but we don't know what specific parts of it actually do the trick and what parts are irrelevant or less important. In other words, we know that a generic walk in the desert caused the death but we still don't know what specific part of that walk (or that cause) exactly caused the effect: and in order to obtain this kind of knowledge we need an approach to causality that allows for a detailed description of the causal relation / chain / process / arrow that generates the effect: eg that explains how the effect actually comes / is brought about.

An autopsy, eg a literal digging into the "inner mechanism" that brought the man to death, helps rejecting the poisoning explanation and declare that the man died of thirst. A close examination of the water can might also rule that the water would have run out of the tank before the poison would properly diffuse. A closer examination of an intervention that is proven to have an effect on a variable might shed light on what components, what activities were actually essential and why and what were not.

This approach is important depending on the use we want to make of **impact evaluation**. If we need it to justify public spending, or for accountability purposes, then a simple link (or an unopened black box) might suffice. But if we need it to improve response, or to extend it to other areas, then we need to know the details of the actual working gears: how can we improve the mechanism or adapt it to / make it work in other areas? What gears are actually important / relevant? What supporting factors are needed in order for the gears to work? Will these supporting factors be present in other areas?

Pawson provides an account of the generative approach as the one enquiring about the "how": how the causal association comes about. "[Within the generative causality framework] the causal explanation provides an account of why the regularity turns out as it does. The causal explanation, in other words, is not a matter of one element (X), or a combination of elements (X1.X2) asserting influence on another (Y), rather it is the association as a whole that is explained. Accordingly, Figure 3 removes the ubiquitous causal arrow and replaces it with the dumbbell shape representing the tie or link between a set of variables. It becomes the causal arrow. And what it explains is the uniformity under investigation. It explains how the outcome pattern is generated. It explains the overall shape of the outcome pattern and thus the causal arrow penetrates, as it where, to its core." (Pawson 2007)

Figure 3: Generative causation

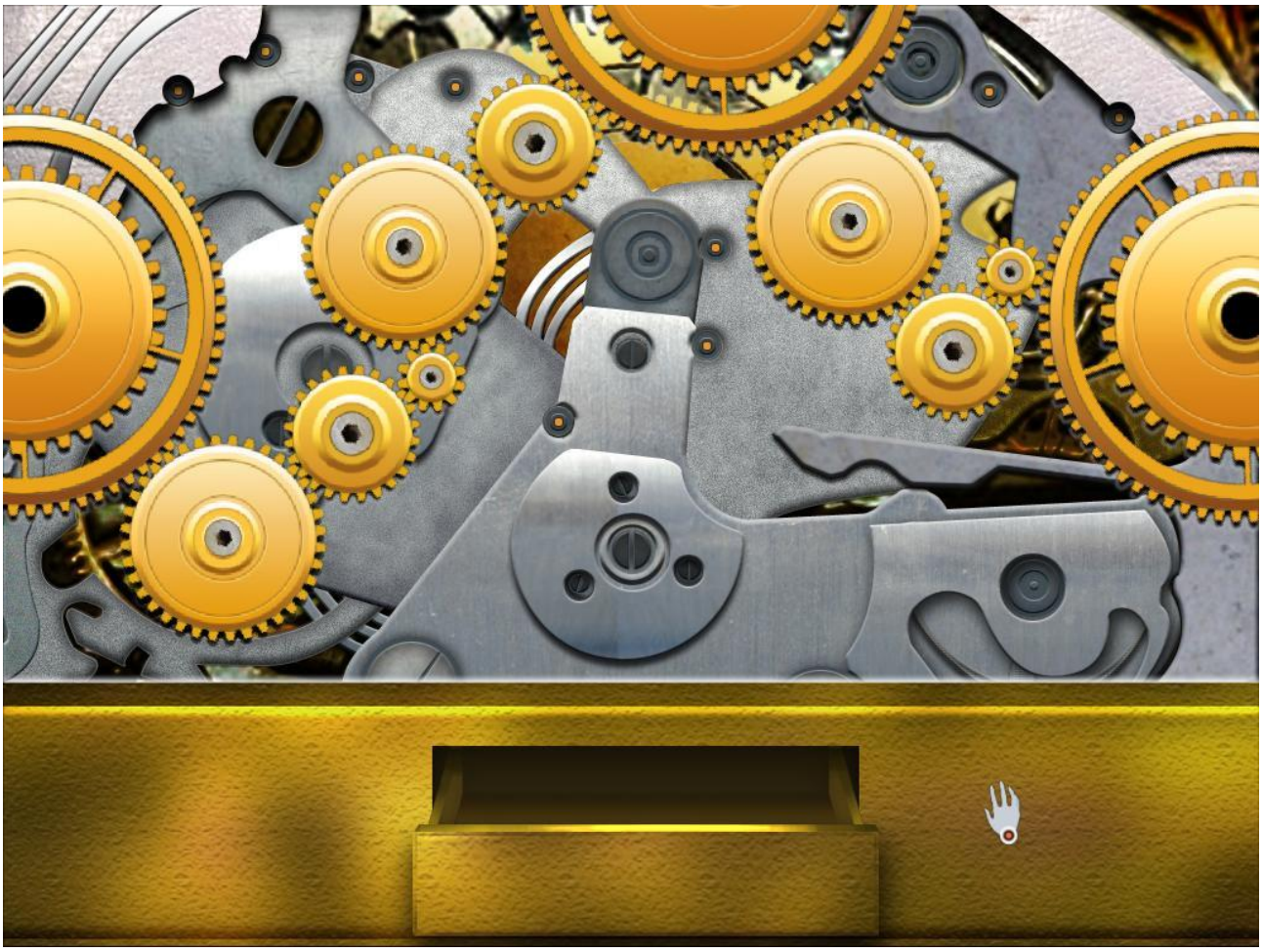


Other theorists claim that mechanisms are “entities and activities organized such that they are productive of regular changes” (Machamber, Darden and Craver 2000) and “causal chains” that provide a more fine-grained, a more satisfactory explanation than a black box regularity (Elster 1998).

3.2.1 How causation is claimed: digging deep

Glennan (1996) stresses the regularity property of mechanisms, arguing that “two events are causally connected when and only when there is a mechanism connecting them” and “the necessity that distinguishes connections from accidental conjunctions is to be understood as deriving from an underlying mechanism”. At the same time, other authors write that mechanisms “misbehave” (Pawson 2007): we don’t always know which one will be activated when; even though when they are, we can recognize it after the fact (Elster 1998.

These apparently contrasting accounts can be reconciled by the fact that mechanisms are “meso” entities that come in different “sizes” and belong to different levels and layers: micro-mechanisms are small parts of higher-level, “bigger” mechanisms (macro-mechanisms or systems). The law-like regularity mentioned by Glennan thus refers to higher-level mechanisms, which are often described in terms of and represented by a causal chain (or intersections of causal chains): or an assemblage of lower-level mechanisms, that have roles in different chains, and might play different roles in each, like gears of different shapes and sizes in a clock or in a manufacturing machine. This is why micro-mechanisms can contribute to different outcomes, depending on the chain they are operating in and the place they occupy in the chain. But at the same time, the whole chain – or an even bigger group of intersections of various causal chains – produces the result in a more “lawlike” than accidental way (eg long pathways and complex systems).

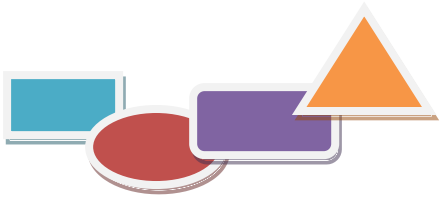


Cartwright claims that, like experiments, causal chains "clinch" conclusions: eg they are proper laws. Although lawlike statements are rarer in the social sciences than in the natural sciences, a fairly articulate and comprehensive (social science) causal chain is able to reject an enormous amount of alternative explanations for an effect, many more than a successionist inference. In fact, while successionism eliminates single causes one by one, a fine-grained explanation containing the same number of single causes rejects a higher number of alternative explanations, because chains with single causes variously combined are different entities even when they include the same causes. Order of conditions and relations between conditions also matter. Single conditions are not linked directly and independently to the effect: rather the whole assemblage is. Therefore, in the case of generativism, by "rival explanations" we do not mean rival single causes or rival combinations, but rather "rival chains" (Campbell in Yin 2003).

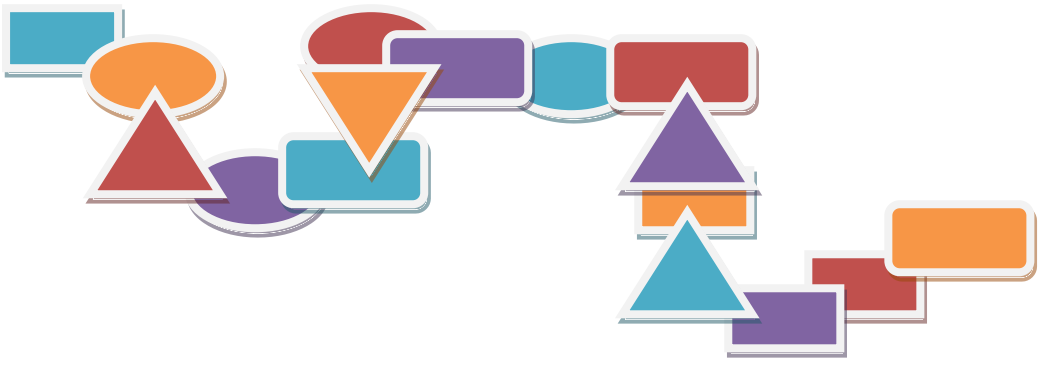
3.2.2 Quality of inference

How fine-grained does the explanation need to be? When is it fine-grained enough? The fit-to-purpose criterion holds here. Even in hard sciences, there is no model or theory depicting absolute truth – theories and paradigms hold until new, better theories and paradigms that unlock a wider, better range of applications are discovered. Usually, the more fine-grained the explanation is, the more in-depth the understanding of the micro-mechanisms that are present in different parts of the chain/system, including their relations; and the greater the possibilities of generalizing the findings to situations presenting similar conditions.

WEAK INFERENCE



STRONG INFERENCE



Determining the causal contribution of several elements helps bring into focus the contribution of the intervention – by subtraction. The more we know about how many factors influence the outcome, the less

we have left to learn about "all other possible elements / alternative explanations". In a reality with a finite number of elements / explanations, considering an ever increasing number of elements eventually leads to rejecting all possible alternative explanations, and thus leads to certainty of inference according to the method of agreement.

3.2.3 Mechanisms have parts: component causes and complete mechanisms

Because they must provide an account of how causality works, mechanisms are often complex and / or complicated objects with several parts. Agent-based modeling illustrates the complexity of macro mechanisms emerging from a high number of micro-mechanisms being activated at the agents' level (Gilbert and Troitzsch 2005, Miller and Page 2007). Rothman and Greenland (2005) avoid the micro-macro distinction and call micro-mechanisms "single component causes". Each component cause is a necessary part of a complete causal mechanism that is sufficient for the effect. The same effect can be achieved also by other causal mechanisms, which may or may not have some component causes in common.

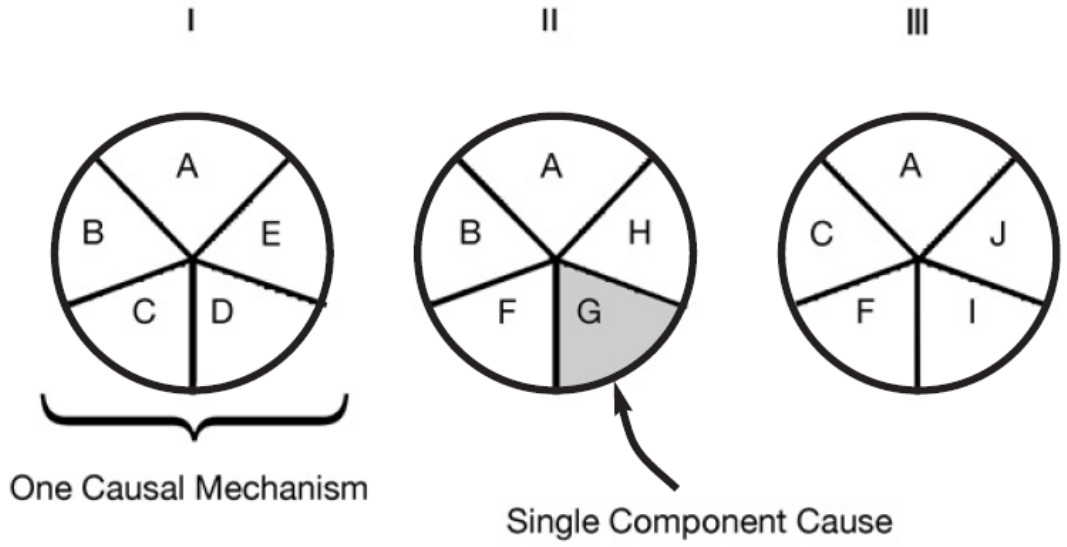


FIGURE 1—Three sufficient causes of disease.

The above – with its considerations of necessity and sufficiency ("completeness") – is reminiscent of configurational causation: the difference is that in this case the linkages among the conditions that constitute the combinations are expressly described and understood, and configurations are not allowed to be mere sets of simultaneously present causes.

Concluding remarks

This review has attempted to highlight the strengths and weaknesses of different models of causality and causal inference. In theory, the most refined way to understand causality is the causal mechanism. The problem with the causal mechanism is that it might be difficult to untangle in all its intricacies, particularly in a complex systems framework. While we try to identify / describe all the steps in a causal chain, or

describe the functioning of a micro-macro interaction, agent-based mechanism producing emergent macro outcomes, regularities and counterfactuals (eg Mill's methods) help ascertaining relevant correlations / associations that, although rough, might signal mechanistic causal relations.

Configurational methods refine the association between cause and effect by "peeking in" in the black box and spotting the presence / absence of a number of elements / conditions, preparing the ground for mechanism identification and theory development. This is an iterative process, with configurations in turn displaying some characteristics of mechanisms and serving as instruments for generalizing / testing theory across cases with limited diversity. Successionism is then eventually able to generalize / test the theory across a high number of similar cases, when theory is enough advanced to allow for the reduction of cases to sets of independent variables.

In conclusion, lessons can be learned from all models of causation and causal inference, and might be summarized in 3 criteria that a causal claim / link needs to meet. Whenever possible, a causal claim / link needs to:

1. Include observation of simultaneous presence of cause and effect. Simultaneous presence might mean:
    a. constant conjunction through several cases / events, while other elements / potential causes vary;
    b. double observation of conjunction cause-effect and conjunction no cause-no effect in two identical situations differing only in cause and effect.
2. Include an analysis of necessity and sufficiency of the cause with regard to the effect:
    a. under what conditions / in what combinations with other causes the effect is produced / unlocked / triggered
    b. When the cause is absent, what other causes / combinations of causes produce the effect
3. Include a detailed, theoretically-informed and theoretically-consistent description of the causal relation / force / chain / "arrow", or how the cause produces the effect.

The following table summarizes several of the arguments presented in this paper.

	Mere Co-Presence			**Active Causality**	
	Of independent causes and effect		**Of causal packages**	**Accidental**	**Regular**
	Regularity	**Counterfactuals**	**Configurations**	**Manipulation**	**Mechanisms**
Major problem solved	Lawlike generalization	Single-cause validity	necessity / sufficiency of packages and single causes within packages	Direction	Pre-emption and pairing
Inference Design	Agreement	Difference	Systematic Comparison	Experiments	In-depth examination (microscope)
Inference Method	Regression, statistical modeling, observational studies	Natural EXPs, Quasi-EXPs w/ control, observational studies	Qualitative Comparative Analysis (QCA)	RCTs, Laboratory experiments in controlled settings	Discovery, construction and refinement of substantive theory
The causal	Simultaneous	Simultaneous	Conditions	What is the	Description of

process or “arrow”	presence of “ends”	presence of “ends”	included between ends	head / nock (direction)	shaft (including head and nock)