a0005 Counterfactual Reasoning, Qualitative: Philosophical Aspects

Stathis Psillos, University of Athens, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Hájek, volume 4, pp. 2872–2874, © 2001, Elsevier Ltd.

Abstract

abspara0010 This article reviews the two major approaches to counterfactual conditionals: the metalinguistic or 'support' approach and the possible worlds approach. It identifies the major problems they face and explores the idea that the core idea behind counterfactual reasoning is to assert that there are not good inductive reasons to affirm simultaneously a generalization and the physical possibility of an exception to it. It also examines the role of counterfactuals in causal inference, causation, and laws of nature.

p0015 Subjunctive conditionals or counterfactual conditionals are probably as old as language itself since they give speakers the means to talk about what would or might happen or have happened if certain things were to happen or had happened. In ordinary language, they have the form:

u0010 If *x* were (not) the case, then *y* would (not) be the case,
u0015 or
u0020 If *x* had (not) been the case, then *y* would (not) have been the case.

p0035 Subjunctive conditionals leave open the possibility of the realization of whatever is expressed in the antecedent, for example, if John were to come to the party, Mary would not go. Counterfactual (or 'contrary-to-fact') conditionals are such that the antecedent is false; the state-of-affairs expressed in it has not actually obtained, for example, if John had gone to the party, Mary would not have gone. (Here it is an implicit assumption that the actual course of events is that John did *not* go to the party.) Both kinds of conditional contrast to indicative conditionals of the form: *if x is the case, then y is the case*. Although there are differences between them, we will not be detained by them, and instead concentrate on counterfactual conditionals. (From now on, we will follow customary usage and use □→ to express the counterfactual 'if…, then…'.)

p0040 Counterfactuals fail a number of principles that indicative conditionals satisfy. Most important, they are nonmonotonic, that is they fail the principle of strengthening of the antecedent:

$$X\Box\!\!\rightarrow Y \text{ does not entail } X\&Z\Box\!\!\rightarrow Y.$$

p0045 *Example*: If John had been poisoned, he would have died. This does not entail: if John had been poisoned and taken an antidote, he would have died.

$$X\Box\!\!\rightarrow Y \text{ and } Y\Box\!\!\rightarrow Z \text{ does not entail } X\Box\!\!\rightarrow Z.$$

u0025 Transitivity:

p0055 *Example*: If John had gone to the market, he would have taken the bus; if John had taken the bus, then he would have gone to his office. These two do not entail: if John had gone to the market, then he would have gone to his office.

$$X\Box\!\!\rightarrow Y \text{ does not entail not} - Y\Box\!\!\rightarrow \text{not} - X.$$

u0030 Contraposition:

p0065 *Example*: If John had lived in a Euro-zone country, he would have used Euros. This is not equivalent to: If John had used Euros, he would have lived in a Euro-zone country.

p0070 If we assume that classical semantics apply to indicative conditionals (the indicative conditional is true if either the antecedent is false or the consequent it true), trying to apply classical semantics to counterfactuals leads to their trivialization: Given the actual falsity of the antecedent of a counterfactual, both the counterfactual with the actual consequent and the counterfactual with the negation of the actual consequent end up being true.

p0075 *Example*: Given that the vase was not struck with a hammer, both of the following two conditionals (treated as material conditionals) are true:

u0035 If this vase had been struck with a hammer, it would have broken, and
u0045 If this vase had been struck with the hammer, it would not have broken.

p0095 The failure of the three principles and this unwanted consequence is a *reductio~~n~~* of the view that classical semantics appl~~y~~ to counterfactuals. But then, what is the right semantics for counterfactuals? What are the truth-conditions of a counterfactual conditional? Or, at least, what are their assertibility conditions? This problem came under sharp focus in the 1940s, when philosophers started to realize that the concept of counterfactual conditionals is instrumental for the explication and understanding of a number of other philosophical concepts. As Nelson Goodman put it in one of the first papers to deal with this issue, "(…) if we lack the means of interpreting counterfactual conditionals, we can hardly claim to have any adequate philosophy of science" (1947, 113).

p0100 Note that in assessing a counterfactual assertion $X\Box\!\!\rightarrow Y$, we should replace, as it were, the actual nonoccurrence of X with the supposition that X has occurred. But given that the laws of nature and the actual course of events led to non-X, in supposing the actual occurrence of X, we need to make counterfactual suppositions concerning either the laws or the actual course of events, such that X actually occurred. In particular, we have to assume that either some laws were broken (so that X did happen after all) or that some actual particular matters of fact did not occur. Hence, in specifying the semantics of counterfactuals, we have to take into account considerations concerning the laws of nature and other particular matters of fact ~~before~~ the conditions specified in the antecedent of the counterfactuals.

p0105 There are two major views concerning the semantics of counterfactuals, the first being introduced by Goodman himself, whereas the second was developed by Robert Stalnaker and David Lewis (but introduced by [bib16] William Todd in 1964). Let us examine them in turn.

s0005 The Metalinguistic or 'Support' View

p0110 On the first major view, known as 'support view' or 'metalinguistic view', a counterfactual conditional $X\Box\!\to Y$ is an elliptic or telescoped argument (or a linguistic construction *about* an argument) such that the antecedent X (taken in its indicative form) together with suitable auxiliary premises entail~~s~~ the consequent (taken in its indicative form). Hence, $X\Box\!\to Y$ should not be taken to be a statement at all; its assertoric content is captured by the following argument type:

$$X\&S\&L \text{ (materially) imply } Y,$$

where L are statements capturing laws of nature and S are singular statements capturing background or collateral conditions that should be 'cotenable' with the antecedent X and express necessary conditions for the consequent to follow.

p0115 *Example*: If this match had been struck (X), it would have lit (Y). For a struck match to light, it is necessary that the match is well made; that it is dry; that there is oxygen; and so on. But ~~even~~ these conditions (collectively designated by S) are not sufficient for the lighting of the match; various laws are required (collectively designated by L). Hence, in asserting the counterfactual 'if this match had been struck, it would have lit', we are committed to the truth of the various statements that describe the laws and the relevant background conditions.

p0120 The first general problem with this view concerns the characterization of the relevance relation when it comes to the background–collateral conditions. It cannot be too permissive. If we allowed all true statements to be relevant to the argument, the falsity of the antecedent X (which is *actually* false) would be relevant, too; however, the counterfactual then would be trivially true. It cannot be too restrictive either. The consequent of the counterfactual is false as well. Hence, not-Y is the case. It is not hard to see that given ($X\&S\&L\to Y$) and not-~~X~~ and L, it follows (by obvious steps) that $X\to$not-S. (Assuming, for simplicity that S is 'the match is dry', the conclusion would be as follows: If the match is struck, it will not be dry!) The point then is that only those background or collateral conditions that are 'cotenable' with the antecedent should be admitted. But which are they? Those conditions S, which are such that if X had been true, S would have been true, too. This is a counterfactual assertion and Goodman thought that this kind of circularity impairs the metalinguistic analysis of counterfactuals.

p0125 The second general problem with this view concerns the characterization of laws, which are indispensable for the *connection* between the antecedent and the consequent of a counterfactual. The key thought here is that some generalizations, although true, are unable to 'support' counterfactuals because they are accidental. Example: Compare the following two counterfactuals:

o0010 A. If x had been a golden sphere, its diameter would not have been more than one mile long.

o0015 B. If x had been a plutonium sphere, its diameter would not have been more than one mile long.

p0140 (A) is false, while (B) is true – we rightly suppose. And this is because there is a law of nature backing up (B), while the generalization related to (A) is merely accidental (intuitively: if we have had enough gold, we could build a sphere of it with the required diameter; not so with plutonium). So the general statements L that are part of the premises of the argument whose telescopic form is $X\Box\!\to Y$ must express *laws of nature* and not merely accidentally true generalizations. But how exactly are we to distinguish between laws and accidents? If we felt that laws are those generalizations that support counterfactuals, whereas accidents are those that do not (see the previous example), then we would move in a(nother) circle. So there is a need to look for ways to distinguish between laws and accidents that do not rely (in the first instance, at least) on their modal force (at least when expressed in their support of counterfactuals). When Goodman brought this problem to the attention of philosophers, the prevailing view of laws was that they are simply regularities (cf [bib2] Chisholm, 1946); hence, the distinction between laws and accidents (which are regularities too) was taken to be mostly an 'honorific' distinction that is captured by the different epistemic attitudes we have toward them. For instance, laws are those regularities that are projected to the future or that are conformable by their instances. [bib14] Sellars (1956: 268), however, pointed out that even if laws are taken to be regularities, they are those regularities that are characterized by 'neck-sticking-out-ness', where this characteristic is captured in the subjunctive mood: 'If this were an A-situation, it would be accompanied by a B-situation'. The counterfactual content of a law then is seen as a 'contextual implication' of a law-statement.

This idea of 'contextual implication' is captured by the supposition view of counterfactuals, which is akin to (although interestingly different from) the metalinguistic view, as this was developed by [bib9] John Mackie (1973). According to this view, to assert something like $X\Box\!\to Y$ is to assert Y *within the scope of the supposition that X*. In other words, we suppose X and then we envisage various possibilities and consequences. This account brings to light the contextuality of counterfactual conditionals, which is not resolvable without some degree of arbitrariness: X did not happen; supposing that X did happen, what else do we have to assume or suppose? What features of the background (including laws and particular matters of fact) should we retain or change? There is no uniquely determined answer to this question, although contextual matters (including a fuller specification of the antecedent of the conditional) might (and as a rule do) help us. *Example*: Is the counterfactual: 'If I had let go of this stone, it would have fallen to the ground' true or false (or assertible or not assertible)? It depends on the context. There are certain conversational contexts in which it would be false to assert it, for example, if this were a precious stone and the owner was very careful with it, so if the stone were to be let go, she would have caught it in midair. *Another example*: Consider the following pair of counterfactuals: 'If Julius Caesar had been in charge of United Nations Forces during the Korean War, then he would have used nuclear weapons' and 'If Julius Caesar had been in charge of United Nations Forces during the Korean War, then he would have used catapults'. Only contextual assumptions can tell us which one, if any, and in what context, is true (or assertible). p0145

The supposition view takes it that counterfactuals are *not* truths about possible words but are ways to express an attitude p0150 toward a possible state of affairs made within the scope of a supposition. The cotenability problem is solved by a legitimate weakening of the cotenability condition: The cotenable premises are taken to be those that are *thought* to be cotenable in a certain conversational context. But how, within this view, can it be explained that laws support counterfactuals while accidents do not? The difference is not in the content of a statement expressing a causal law as opposed to the content of an accidentally true generalization. Rather, the difference is in the circumstances under which it is legitimate (or acceptable) to combine the supposition that X is the case (i.e., the antecedent of the counterfactual X□→Y) with the law L as premises of the relevant argument whose conclusion is Y. Suppose that the sole ground for believing the law L (e.g., All Fs are G) is an enumeration of *all* actual instances (Fai&Gai) of L. Then adding the supposition X, that is, that a *further* a is F, removes the ground for accepting L. We can no longer draw the conclusion that this further a is G. Hence, we cannot assert the counterfactual X□→Y. More generally, if the reasons for accepting L survive placing L within the scope of *the supposition that there are further instances of the law's subject term*, then we can say that the law supports the relevant counterfactual conditional. According to Mackie the required reasons are ordinary inductive reasons, that is, good inductive evidence for the law. Good inductive evidence, in other words, is evidence for the 'neck-sticking-out-ness' of the law. As Mackie (1974: 203) noted, the evidence plays a *double role*. It first establishes inductively a generalization. But then, "it continues to operate separately in making it reasonable to assert the counterfactual conditionals which look like an extension of the law into merely possible worlds" (Mackie 1974).

p0155 An interesting related thought comes from Julius Weinberg (1951) who claimed the following: A counterfactual X□→Y is not best seen as the indicative statement (the statement of a generalization) X → Y plus some further antecedent conditions (including that X did not actually happen), but rather as asserting something about the evidence there is for X→Y, that is, that there is evidence for, and no evidence against, the gereralization: for all X (X→Y). Hence, the additional strength a counterfactual is supposed to have over the corresponding generalization is captured by the evidence there is for the generalization.

s0010 The Possible Worlds View

p0160 Taking literally the view that counterfactuals are used in contemplating *possibilities*, the second major view of the semantics of counterfactuals appeals to possible worlds. In first suggesting this view, Todd (1964: 107) noted that when we allow for the possibility that the antecedent of a counterfactual be true, we are "hypothetically substituting a different world for the actual one". On this view, the core meaning of a counterfactual X□→Y is (roughly): In the possible (but not actual) world where X, Y too.

p0165 A possible world is a way the world might be or might have been. For instance, it is possible that gold is not yellow, or that planets describe circular orbits, or that birds do not fly, or that beer does not need yeast to brew. But are there really possible worlds? There are three views here. The *first* is that talk of possible worlds is a mere *facon de parler*, although useful when it comes to assessing counterfactuals (cf Mackie, 1974: 199). (I take it that an extension of this view is that possible worlds are useful *fictions*.) The *second* is 'extreme realism', according to which the way the world *actually* is, is one among the many ways the world could be; hence, the actual world is one among the many possible worlds, the latter being no less real than the actual. The chief advocate of this view was David Lewis (1973). The *third* view is 'abstract realism', according to which possible worlds are maximally consistent sets of propositions: total ways things might be. A 'possible world' then is fit to represent a concrete reality, but only one of them actually represents anything, that is, the actual world (cf Bennett, 2003).

Stalnaker (1968) developed the core meaning of counterfactuals as follows: p0170

Consider a possible world W in which X is true but otherwise is similar to the actual world @. X□→Y is true if Y is true in W. p0175 AU1

The similarity relation among worlds (a selection function, as Stalnaker put it) is an ordering of possible worlds with respect to their resemblance to the actual world. p0180

Calling an X-world a possible world in which X hold, counterfactuals might be taken to be strict conditionals of the following form: p0185

X□→Y is true in a world W if Y is true in all X-worlds such-that _____ where the blank is filled by a general condition that X-words should satisfy. Hence, whatever goes into the blank _____ places a restriction on the admissible (or accessible) possible worlds. This idea would model counterfactuals along the lines of strict conditionals of the form: u0050

it is physically necessary that _____ u0055
or u0060
it is logically necessary that _____ u0065

where the first restriction is to all worlds with the same laws as the actual, whereas the second 'restriction' would be to all possible worlds *simpliciter*.

But this analysis cannot be correct. There is no set of possible worlds W such that X→Y throughout W (this is another way to state the fact that counterfactuals are nonmonotonic). So Lewis (1973) suggested that counterfactuals X□→Y are *variably strict conditionals*: each of them is a strict conditional, that is, every X-world *of a certain sort* is a Y-world, but the relevant set of worlds varies with different conditionals. p0210

Like Stalnaker, Lewis took it that worlds are ordered in terms of similarity, or closeness to the actual world. According to this *primitive* notion of 'comparative overall similarity': "we may say that one world is closer to actuality than another if the first resembles our actual world more than the second does, taking account of all the respects of similarity and difference and balancing them off against one another" (1986: 163). p0215

But unlike Stalnaker, Lewis took it that in assessing the counterfactual X□→Y, it does not make good sense to talk about *the* closest-to-actual possible X-world. It is not just that there might be more than one closest-to-the actual possible worlds. It is mainly that there might not be even one rightly p0220 deemed *the* closest (even in a limiting sense). Hence, according to Lewis's view:

p0225 $X \Box\!\!\rightarrow Y$ is true at a world W if some (accessible) X-world in which Y holds is closer to W than any X-worlds that Y does not hold.

p0230 For instance, take the counterfactual that if this pen had been left unsupported (X), it would have fallen to the floor (Y). Neither X nor Y are true of the actual world. The pen was never removed from the table, and it did not fall to the floor. Take all X-worlds. The counterfactual $X \Box\!\!\rightarrow Y$ is true (in @) if the X-worlds in which Y is true (i.e., the pen is left unsupported and falls to the floor) are closer to @ than any of the X-worlds in which Y is false (i.e., the pen is left unsupported but does not fall to the ground, e.g., it stays still in midair). As Lewis (1986, 164) put it, "[A] counterfactual (...) is true if it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent."

p0235 The key idea behind the possible-world semantics is that in specifying the truth-conditions of a counterfactual conditional, we should imagine a state of affairs in which X obtains and which is such that all *else is pretty much as they actually were*. But as noted already, this is not quite possible. In the possible world in which X did happen, many other things (including the laws) were different from the actual world @ in which X did not occur. Can we find comfort in the notion of comparative similarity? Now, although 'comparative overall similarity' is not defined strictly, a lot can be said of it. Notably, it imposes a weak ordering on the set of possible worlds that are accessible from @, that is, the relation of comparative similarity is connected and transitive. (It also imposes a centering assumption: @ is closer to itself than any other world is to it.) More important, however, similarity is clearly not one dimensional, but rather it is the result of many component similarities. Lewis (1986: 47–48) ranked possible worlds according to the following dimensions of similarity (put in order of importance):

- u0070 Avoid big, widespread violations of the laws of nature of the actual world (very important).
- u0075 Maximize the spatiotemporal perfect match of particular matters of fact.
- u0080 Avoid small, localized violations of the laws of nature of the actual world.
- u0085 Secure approximate similarity of particular matters of fact (not at all important).

p0260 So, a world W_1 that has the same laws of nature as the actual world @ is closer to @ than a world W_2 that has different laws. But insofar as there is exact similarity of particular facts in large spatiotemporal regions between @ and a world W_3, Lewis allows that W_3 is close to @ even if some of the laws that hold in @ are violated in W_3.

p0265 All this implies that there is quite a lot of vagueness in the notion of overall comparative similarity, which accounts for the fact that counterfactuals themselves are vague, at least in the sense that it is a contextual matter as to what to keep fixed and what to change when we assert a counterfactual conditional. A more serious worry relates to the issue of the motivation behind the foregoing ranking of dimensions of similarity among worlds. It has been observed by many that Lewis's initial theory yielded the wrong truth-values for a type of counterfactual conditional that can be schematized as follows:

$$X \Box\!\!\rightarrow \text{BIG DIFFERENCE}.$$

For instance: p0270

(C) If the president had pressed the button, a nuclear war would have ensued. p0275

p0280 Intuitively, (C) is true. But on Lewis's initial account, it would be false. For a possible world W_1 in which the president did press the button and a nuclear war did erupt is more distant from (because more dissimilar to) the actual world than a world W_2 in which the president did press the button but, somehow, a nuclear war did *not* follow. Addressing this worry, Lewis noted that intuitive judgments of the truth and falsity of counterfactuals are ~~before~~ the similarity relation that is required for the semantics of counterfactuals; hence, the similarity relation should be such that it tallies with the right intuitive judgments concerning counterfactuals. The similarity ranking is meant to solve this problem. To see how the foregoing counterfactual is indeed true, Lewis invited us to consider the following: Take a world W_1 in which nothing extraordinary happened between the president's pressing the button and the activation of the nuclear missiles. In W_1 the nuclear war did erupt. Take, now, a world W_2 in which the president did press the button, but the nuclear war did *not* follow. For this to happen, many miracles would need to take place (or, to put it in a different way, a really *big* miracle would have to occur). For all the many and tiny traces of the button pushing would have to be wiped out. Hence, appearances to the contrary, W_2 would be more distant from (because more dissimilar to) actuality @ than W_1. The *big* violation of laws of nature in W_2 is outweighed by the maximization of the perfect spatiotemporal match of particular matters of fact between W_1 and @. So, with the help of the refined criteria of similarity among possible worlds, the president counterfactual becomes true. Still, one may follow Horwich (1987: 171–172) in wondering how psychologically plausible Lewis's theory becomes: The similarity criteria are so tailored that the right counterfactuals become true, but they have little to do with our pretheoretical understanding of judgments of similarity.

p0285 As noted in relation to the 'support' view, an adequate theory of counterfactuals has to jump two hurdles. The first relates to cotenability. Lewis solved this problem by taking it that some conditions S are cotenable with X (the antecedent of the counterfactual $X \Box\!\!\rightarrow Y$) if some X-world is closer to the actual world than any not-S world. The second hurdle relates to the distinction between laws and accidents. Here, the possible world approach is on safe ground, although the ground can support any decent theory of counterfactuals. David Lewis (1973) revamped a long tradition that goes back to John Stuart Mill, via Frank Ramsey, according to which the regularities that constitute the laws of nature are those that are expressed by the axioms and theorems of an ideal deductive system of our knowledge of the world, and in particular, of a deductive system that strikes the *best* balance between simplicity and strength. Simplicity is required because it disallows extraneous elements from the system of laws. Strength is required because the deductive system should be as informative as possible about the laws that hold in the world. Whatever regularity is not part of this *best* *system*, it is merely accidental. The gist of this approach is that no regularity, taken in isolation, can be deemed a law of nature. The regularities that constitute laws of nature are determined in a kind of holistic fashion by being parts of a structure. An advantage of this approach is that it can sustain, in a noncircular way, the view that laws can support counterfactuals. For, it identifies laws *independently* of their ability to support counterfactuals.

p0290 A key objection to the possible world approach to counterfactuals is that counterfactual conditionals are not purely objective; an irremediably subjective elements enters into the judgment of similarity (and arguably, to the distinction between laws and accidents). Not only are the truth-conditions of counterfactuals "a highly volatile matter" as Lewis (1973: 92) himself noted, but also which counterfactuals are true turns out to depend on various partly nonobjective judgments concerning similarity weights and conversational contexts. This objection, however, might not be as fatal as it first seems precisely because counterfactual conditionals should not be taken to be pointers to necessary connections, powers, and the like but rather (in either of the two theories we have examined) summaries of attitudes we have toward statements that are supposed to express a *connection* between a hypothetical antecedent and a consequent. To exploit an idea of Sellars's, the core idea behind counterfactual reasoning is to assert that there are not good inductive reasons to affirm simultaneously a generalization and the physical possibility of an exception to it.

s0015 Counterfactuals and Evidence

p0295 Whichever way counterfactuals are treated, a particularly acute problem seems to arise, especially for empiricists: How are they connected with the available evidence, which can be gathered only in the actual world? To put the point differently, there cannot be an empirical test for a counterfactual by realizing its antecedent. The supporters of the 'metalinguistic view', especially Mackie, avoid this problem by taking the counterfactual claim to be licensed by good inductive evidence. The advocates of the possible worlds approach pin their hopes on the role laws play in the truth-conditions of the counterfactuals. In particular, the thought is that although counterfactuals are about other possible worlds, these are *very much like* the actual world; hence, the evidence gathered in the actual world can be evidence for the truth of a counterfactual, by supporting the laws that back a similarity claim between the actual world and the world in which the antecedent of a counterfactual obtains.

s0020 Rubin's and Holland's Model

p0300 A promising model of counterfactuals, which aims to offer empirical test-conditions for them, has been developed by Donald Rubin (1978) and Paul Holland (1986) and has attracted increasing interest among statisticians and social scientists. This model focuses on the discovery of the effects of causes. Suppose, to use a simple example, we want to find out whether taking an aspirin makes a difference to a *specific* subject's relief from headache. We would like to give a certain subject u an aspirin to see what happens to the headache episode – let's call the result Y. Ideally, we also would like, *at the same time*, to withhold giving aspirin to the very same subject u, to see what happens to the headache episode – let's call this result Y'. The difference, if any, between Y and Y' naturally would be considered the causal effect of aspirin-taking on the headache episode of subject u. But this kind of experiment is impossible: The experimenter cannot give and *not* give an aspirin to the *same* subject u at the *same* time. Rubin's and Holland's main idea is that an appeal to counterfactuals allows us to make an inference about the causal effect.

p0305 Let's consider a population U of individuals, or units, $u \in U$. In a typical experiment, the experimenter applies one treatment, say i, out of a set of possible treatments T, to each unit u and observes the resulting responses Y. The experimental units are chosen and separated into two groups (the experimental group and the control group) by randomization. To simplify matters, let the treatment set T consist of two possible actions (treatment – t, and control – c). For instance, t may be taking the aspirin and c may be taking a placebo. Let Y also consist of two possible responses, for example, headache relief – Y_t, and headache persistence – Y_c. Although it is crucial that each unit is potentially exposable to any one of the treatments, to each unit u just one treatment is *actually* given, that is, either t or c. Similarly, for each unit u, there is just one response that actually is observed, that is, either $Y_t(u) = Y(t, u)$ or $Y_c(u) = Y(c, u)$. Rubin's model defines the two responses in subjunctive–counterfactual terms. $Y(t, u)$ is the value of the response that would be observed if the unit u were exposed to treatment t and $Y(c, u)$ is the value that would be observed *on the same unit* u if it were exposed to c. A key assumption of Rubin's model is that both values $Y(t, u)$ are $Y(c, u)$ are well defined and determined. In particular, it is assumed that even if subject u actually is given treatment t and has response $Y(t, u)$, there is still a fact of the matter about what the subject's u response would have been, had she been given treatment c. The task is to figure out the *individual causal effect*, that is the difference AU2

$$\tau(u) = Y(t, u) - Y(c, u), \qquad [1]$$

which measures the effect of treatment t on u, relative to treatment c.

p0310 In each particular experiment, either $Y(t, u)$ or $Y(c, u)$ (but not both) ceases to be counterfactual. Yet, given that one of $Y(t, u)$ and $Y(c, u)$ becomes testable, the other *has to* be untestable. Holland has called a situation such as this "the *fundamental problem of causal inference*". Does it follow that figuring out eqn [1] is impossible?

p0315 Suppose that we give treatment t to u and we observe $Y(t, u)$. The question then is how could we possibly figure out the counterfactual value of $Y(c, u)$? According to the present model, when *certain assumptions are in place*, there are ways to assess counterfactuals such as the above. Here is how:

p0320 Given that unit u got treatment t, we may try treatment c to a *different* unit u', which is very much like u, except that it was given treatment c instead. That is, instead of testing the counterfactual conditional $Y(c, u)$, which is impossible, we test the indicative conditional $Y(c, u')$ – the response of unit u' if she is given treatment c – and claim that this tells *indirectly* what the value of $Y(c, u)$ is. For this move to be plausible at all, we need an assumption of *unit homogeneity*: that u and u' are so similar that the actual response of u' to treatment c is the same as the response that unit u would have to treatment c. Under this assumption, we take it that $Y(t,u) = Y(t,u')$ and $Y(c,u) = Y(c,u')$. Then, the individual causal effect can be calculated, since eqn [1] becomes thus:

$$\tau(u) = Y(t,u) - Y(c,u) = Y(t,u) - Y(c,u'). \qquad [2]$$

p0325 Although eqn [1] involves essentially a counterfactual conditional ($Y(c, u)$), eqn [2] does not. Eqn [2] is indeed testable, but the counterfactuals are gone. Instead, eqn [2] has two indicative conditionals, one for unit u who received treatment t and another for unit u' who received treatment c. In a sense, the unit homogeneity assumption renders the counterfactual conditional $Y(c, u)$ not so much a claim about the *specific* unit u but rather a claim about *any* of the homogeneous units. It is because of this fact that the counterfactual is supposed to become testable.

p0330 We might proceed in another way to calculate $\tau(u)$. Instead of giving treatment t to unit u and treatment c to (uniform) unit u', we give treatment c to unit u at time t_1 and treatment t to the *very same unit* u at a later time t_2. This move requires another assumption, that is, *temporal stability* or the constancy of response over time. It also requires an assumption of 'causal transience' to avoid situations like this: The subject's taking a placebo at time t_1 changes some properties of her enough to affect her response to taking an aspirin at a later time t_2. Under these assumptions, we take it that $Y(t_{t_1}, u) = Y(t_{t_2}, u)$ and $Y(c_{t_1}, u) = Y(c_{t_2}, u)$. If this is so, then the individual causal effect can be calculated, because eqn [1] becomes:

$$\tau(u) = Y(t,u) - Y(c,u) = Y(t_{t_2}, u) - Y(c_{t_1}, u'). \qquad [3]$$

p0335 The remarks made about eqn [2] can be repeated about eqn [3], too. Eqn [3] has no counterfactuals and it seems that the content of eqn [1] – which does involve the counterfactual $Y(c, u)$ – *reduces* to the joined content of two indicative conditionals $Y(t_{t_2}, u)$ and $Y(c_{t_1}, u)$ together with the two further assumptions of causal transience and temporal stability.

p0340 The key point then is that the alleged testability of counterfactual conditionals is predicated on the plausibility and success of certain general assumptions noted previously. These assumptions might fail. If, however, there are reasons to believe they do not, that is, if there is *evidence* for the general assumptions, then causal inference seems quite safe. I would suggest that these assumptions are characteristics of *stable causal-nomological structures*. Consider *unit homogeneity*. For it to hold, it must be the case that two units u and u' are alike in all causally relevant respects other than treatment status. If this is so, we can substitute u for u' and vice versa. This simply means that there is a causal law connecting the treatment and its characteristic effect, which holds for all homogeneous units and hence it is independent of the actual unit chosen (or could have been chosen) to test it. In effect, this holds for temporal stability too, since the latter is the temporal version of unit homogeneity.

s0025 Interventionist Counterfactuals

p0345 James Woodward recently has introduced the claim that only counterfactuals that are related to *interventions* can be of help when it comes to assessing their test or assertibility conditions. An intervention gives rise to an 'active counterfactual', that is, to a counterfactual whose antecedent is made true by (hypothetical) interventions.

p0350 Woodward (2003: 3) characterized the appropriate counterfactuals in terms of *experiments*: They "are understood as claims about what would happen if a certain sort of experiment were to be performed".

p0355 Take Ohm's law (that the voltage E of a current is equal to the product of its intensity I times the resistance R of the wire) and consider the following two subjunctives:

1. If the resistance were set to $R = r$ at time t, and the voltage were set to $E = e$ at t, then the intensity I would be $I = e/r$ at t. o0020
2. If the resistance were set to $R = r$ at time t, and the voltage were set to $E = e$ at time t, then the intensity I would be $i^* \neq e/r$ at t. o0025

p0370 According to Woodward, we can perform the experiments at a future time t^* to see whether (1) or (2) are true. If, however, we are interested in finding out what *would* have happened, had we performed the experiment in a past time t (although we never did), Woodward invited us to rely on the 'very good evidence' we have that the behavior of the circuit is stable over time. Given this evidence, we can assume that the *actual* performance of the experiment at a future time t^* is as good for the assessment of (1) and (2) as a *hypothetical* performance of the experiment at the past time t.

p0375 An obvious advantage of this approach is that the truth-conditions of (the right sort of) counterfactual conditionals are not specified by means of an abstract metaphysical theory, as in the possible worlds approach. But there is a residual tension in this view. Woodward (rightly) insisted that counterfactual conditionals have determinate meaning and truth-conditions independent of the actual and hypothetical interventions. So there is a distinction between truth-conditions and test-conditions for a counterfactual. But then there must be a way for counterfactuals to get their truth-conditions fixed independent of their test-conditions. It is not quite clear what this way might be. The unclarity is accentuated by the fact that if there is such a way to specify the truth-conditions of a counterfactual conditional independent of its test-conditions (which are related to hypothetical interventions), then this way will offer truth-conditions to counterfactuals that do not (or might not) have test-conditions at all.

p0380 What if we were to collapse the truth-conditions of counterfactuals to their test-conditions? One can see the prima facie attraction of this move. Because evidence-conditions are specified in terms of actual and hypothetical experiments, the right sort of counterfactuals (the active counterfactuals) *and only those* end up being meaningful and truth-valuable. But there is an important drawback. Recall the subjunctive assertion in (1). On the option presently considered, what makes (1) true is that its evidence-conditions obtain. Under this option, counterfactual conditionals lose, so to speak, their counterfactuality. (1) becomes a shorthand for a future prediction or the evidence that supports the relevant law. If t is a *future* time, (1) gives way to an indicative conditional (a prediction). If t is a past time, then, given that there is good evidence for Ohm's law, all that (1) asserts under the present option is that there has been good evidence for the law.

p0385 In any case, Woodward is keen to keep evidence- and truth-conditions apart. One option would be to tie the truth-conditions of counterfactual conditionals to *laws of nature*. It then is easy to see how the evidence-conditions (i.e., actual and hypothetical experiments) are connected with the truth-conditions of a counterfactual: Actual and hypothetical experiments are *symptoms* for the presence of a law. Be that as it may, it is hard to see how a counterfactual can be assessed without taking into account the evidence there is about laws of nature or other general assumptions.

s0030 Causation and Powers

p0390 Despite the various difficulties we have presented, counterfactual conditionals are a stable part of the philosophical arsenal. Their being modal in character has invited the thought that they can capture a special connection between the antecedent and the consequent – most typically *a causal connection*. According to a popular counterfactual analysis of causation, to say (roughly) that event c causes event e is to say that e is counterfactually dependent on c, that is, that if c had not happened, e would not have happened either. This idea goes back to David Hume, but has been developed by Lewis (1974) into a full-blown theory. The sufficiency part of the definition is straightforward: If two events c and e are actual, and e is counterfactually dependent on c, then c is the cause of e. *Example*: Let c be the actual short circuit and e be the actual fire. If it is the case that if c had not occurred, then e would not have occurred, then the short circuit is the cause of the fire. But causation is transitive, whereas (as we have seen) counterfactual dependence is not. *Example*: Let e' be an effect of the fire e, for example, that the owner of the burnt house got some insurance money. If c causes e and e causes e', then c causes e'. Although the owner's insurance compensation (e') is counterfactually dependent on his house getting fire (e), which, in turn, is counterfactually dependent on the short circuit (c), e' is not counterfactually dependent on c: The owner would have got the insurance compensation (e') even if the short circuit (c) had not occurred, assuming that the fire was caused in some other way. To make counterfactual dependence a necessary condition for causation, Lewis introduced a way to enforce the transitivity of counterfactual dependence: The sequence of events must form a *causal chain*. A sequence of events $<c, e, e', \ldots>$ is a chain of causal (counterfactual) dependence if e causally (counterfactually) depends on c, e' causally (counterfactually) depends on e, and so on.

p0395 This position, intuitively compelling though it may be, faces a number of important difficulties that has led to various ad hoc additions and modifications. (For a discussion of them see my 2002 ~~paper~~.) The key point is that this approach explains how the effect *depends* on the cause without entailing anything as to how the effect is *connected* to the cause.

p0400 Those who think that there is a special kind of *connection* between cause and effect take counterfactual dependence to be a symptom of the presence of a power in the cause to bring about the effect: Causation amounts to a power's producing its manifestation. (This idea goes back to Leibniz who took it that causes are 'producers'). Actually, ascription of dispositions to objects (e.g., fragility or elasticity and the like) has been analyzed in terms of counterfactual (and subjunctive) conditionals. So, to ascribe a disposition F to an object x is to say that if x were to be given stimulus S, the characteristic result would be R. *Example*: x is fragile if x were to be struck, it would break. bib4 Rom Harré and Edward Madden (1975) offered a general analysis of powers along the foregoing lines: x has the power to F if x were subject to stimuli or conditions of an appropriate kinds, then x would do F, "in virtue of its intrinsic nature". So power-ascriptions to objects is analyzed in terms of (1) a specific counterfactual conditional and (2) an unspecified categorical claim about the nature of the object.

p0405 Despite its initial promise, this view faces important counterexamples, most of which point to the claim that the meaning of dispositional ascriptions cannot be captured by counterfactual conditionals. For it is possible either that the antecedent of the counterfactual is realized and the characteristic response does not obtain (e.g., because something else blocks the manifestation of the disposition) or that a disposition exists even if there are no manifestations of it (and hence no relevant counterfactuals to be entailed by it). Hence, the power-based attempts to understand the causal connection should either dissociate causation from counterfactuals or take "being disposed to" as a special (not further reducible) relation (see bib11 Mumford and Anjum, 2011).

Acknowledgment s0035

This research has been cofinanced by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF) – Research Funding Program: THALIS – UOA – APRePoSMA. p0410

See also: 63006; 63007; 63016.

References

Bennett, Jonathan, 2003. A Philosophical Guide to Conditionals. Oxford University Press, Oxford.
Chisholm, Roderick M., 1946. The contrary-to-fact conditionals. Mind 55, 289–307.
Goodman, Nelson, 1947. The problem of counterfactual conditionals. Journal of Philosophy 44, 113–128.
Harre, Rom, Madden, E.H., 1975. Causal Powers: A Theory of Natural Necessity. Blackwell, Oxford.
Holland, Paul, 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–960.
Horwich, Paul, 1987. Asymmetries in Time. MIT Press, Cambridge, MA.
Lewis, David, 1973. Counterfactuals. Blackwell, Oxford.
Lewis, David, 1986. Philosophical Papers, vol. II. Oxford University Press, Oxford.
Mackie, J.L., 1973. Truth, Probability and Paradox. Clarendon Press, Oxford.
Mackie, J.L., 1974. The Cement of the Universe: A Study of Causation. Clarendon Press, Oxford.
Mumford, Stephen, Anjum, Rani Lill, 2011. Getting Causes from Powers. Oxford University Press, Oxford.
Psillos, Stathis, 2002. Causation and Explanation. Acumen, Chesham. AU3
Rubin, Donald B., 1978. Bayesian inference for causal effects: the role of randomization. The Annals of Statistics 6, 34–58.
Sellars, Wilfrid, 1956. Counterfactuals, dispositions, and the causal modalities. In: Feigl, Herbert, et al. (Eds.), Concepts, Theories and the Mind-Body Problem, Minnesota Studies in the Philosophy of Science, vol. II, pp. 225–308.

Stalnaker, Robert, 1968. A theory of conditionals. In: Rescher, N. (Ed.), Studies in Logical Theory. Blackwell, Oxford.

AU4

Todd, William, 1964. Counterfactual conditionals and the presuppositions of induction. Philosophy of Science 31, 101–110.

Weinberg, Julius, 1951. Contrary-to-fact conditionals. Journal of Philosophy 48, 17–22.

Woodward, James, 2003. Making Things Happen: A Theory of Causal Explanation. Oxford University Press, New York.