Mechanistic evidence: Disambiguating the Russo-Williamson Thesis

Phyllis McKay Illari

Draft of March 16, 2011

Abstract

Russo and Williamson claim that establishing causal claims requires mechanistic and difference-making evidence. In this paper, I will argue that Russo and Williamson's formulation of their thesis is multiply ambiguous. I will make three distinctions: mechanistic evidence as *type* vs *object* of evidence; *what* mechanism or mechanisms we want evidence of; and how *much* evidence of a mechanism we require. I will feed these more precise meanings back into the Russo-Williamson Thesis and argue that it is both true and false: two weaker versions of the thesis are worth supporting, while the stronger versions are not. Further, my distinctions are of wider concern because they allow us to make more precise claims about what kinds of evidence are required in particular cases.

1 Introduction

The Russo-Williamson Thesis (RWT) is a thesis concerning causal inference that is generating a good deal of criticism (Weber, 2009; Leuridan and Weber, 2011; Broadbent, 2011; Gillies, 2011; Howick, 2011). Russo and Williamson are concerned with inference *to* a causal structure, and their thesis says that both mechanistic evidence (such as might be got from experiments establishing the existence of biochemical pathways) and difference-making evidence (such as might be got from well-conducted RCTs) is required to establish causal claims in the biomedical sciences. In their own words:

> To establish causal claims, scientists need the mutual support of mechanisms and dependencies. The idea is that probabilistic evidence needs to be accounted for by an underlying mechanism before the causal claim can be established. (Russo and Williamson, 2007, p159.)

I will use the term *difference-making* evidence, rather than *probabilistic* evidence or dependency, in line with both Russo and Williamson (2011), and their original claim that probabilistic evidence is used to show that the cause *makes a difference* to the effect (Russo and Williamson, 2007, p158).

In this paper, I will argue that Russo and Williamson's formulation of the thesis is multiply ambiguous, so that it has a *range* of possible interpretations from the very strong and wildly implausible to the drastically weaker and quite plausible. I will make three distinctions: first, mechanistic evidence as *type* versus *object* of evidence in section 2; secondly, *what* mechanism or mechanisms we want evidence of in section 3; and thirdly versions of how *much* evidence of a mechanism we require in section 4. These distinctions will allow me to generate more precise versions of the RWT and argue in section 5 that while the stronger versions of the thesis should be denied, two of the weaker versions are worth supporting. And on most versions, mechanisms come out as *helpful* in causal inference.

In this paper I will be examining mechanistic evidence and I will say comparatively little about any other kind of evidence. But I do not believe that mechanistic evidence can *replace* good difference-making evidence, such as that gained from well-conducted RCTs. My aim, along with Russo and Williamson, is to examine how good mechanistic evidence can *complement* good difference-making evidence in establishing causal claims. I aim to show that in denying mechanistic evidence *any* role, the Evidence-Based Medicine movement is missing something that, for example, the International Agency for Research in Cancer has got hold of. (See IARC (2006), and also Leuridan and Weber (2011) for constructive criticism.)

I also intend the distinctions I make to assist in theorizing about causal inference, and in its practice, by allowing us to make more precise claims about mechanistic evidence. In section 2.2, I will begin examining how evidence of mechanism can help decide between the three possible explanations of a correlation: causal relation, accident, or confounding. That mechanisms can help with this is an idea that has been around for a while, but we need the distinctions I will introduce to avoid muddle. To anticipate, they will allow us to make precise claims such as: in this case we have direct relatively complete evidence of similar kinds of mechanisms operating in the domain, plus a plausible story about a mechanism linking cause and effect, but no direct evidence of the actual existence of the postulated mechanism linking cause and effect. This will allow a clearer view of the evidential support we have for causal claims.

Finally, I will here focus solely on the roles mechanistic evidence might play in *establishing* the claim that C causes E. Various other claims have been made for the role of mechanisms in the biomedical sciences: causal explanation (Vreese, 2008; Leuridan and Weber, 2011; Russo and Williamson, 2007); assisting with the problem of external validity (Leuridan and Weber, 2011; Russo and Williamson, 2007); assessing the stability of established causal claims (Leuridan and Weber, 2011); and assessing causal effect size in epidemiology (Kincaid, 2011). These are all interesting ideas which further support the need for understanding the use of mechanisms in causal inference, but I do not address them here.

2 First distinction: *method of gathering* or *object* of evidence?

In claiming that establishing causal claims requires mechanistic evidence as well as difference-making evidence, there are two different distinctions that might be meant, corresponding to two different ways of categorising types of evidence.

First Distinction

1. There are two types of evidence-gathering *methods*—mechanistic and difference-making methods—each yielding different types of evidence.

2. There are two *kinds of things* we have evidence *of*: what you have evidence of when you have evidence of a mechanism, and when you have evidence of difference-making, are different.

These are not the same claim. The first categorises evidence by the methods used to *gather* it, just as we categorise, say, x-ray crystallography evidence, electron microscope evidence, biopsy evidence or blood test evidence by the tools or techniques we use to gather the evidence. The second categorises evidence by the kinds of things we have evidence of, the *objects* of evidence, just as we categorise evidence of the structure of DNA, evidence of replication forks, evidence of cancer, and evidence of low blood sugar levels, by their objects. Once these different claims are clear, it can be seen that they are independent: there is no particular reason to assume, in advance of investigation, that any particular evidence-gathering method can only be used to gather evidence of one particular type of thing.

Russo and Williamson do not rule out the first reading, and so lay themselves open to be interpreted as claiming that establishing a causal claim requires an extra type of evidence (mechanistic evidence), with a distinctive class of evidence-gathering methods, in addition to the usual type of evidence (difference-making evidence), gathered using such familiar methods as RCTs and observational studies. There are some things that Russo and Williamson say in the early paper that are ambiguous regarding these two readings:

> two different types of evidence – probabilistic and mechanistic – are at stake when deciding whether or not to accept a causal claim. (Russo and Williamson, 2007, p163).

In this section I will argue that if the RWT is interpreted as claiming that the use of a distinctive class of evidence-gathering methods is required for causal inference, then it is false. Instead, Russo and Williamson should be read as talking about objects of evidence. This is in line with their more careful claims in their latest paper, such as:

> In the following, we will show ... that at each level, causal claims depend on evidence both of difference-making and of mechanisms. (Russo and Williamson, 2011, p10 of manuscript.)

I will argue in section 2.1 that types of evidence are usefully categorised by evidence-gathering methods when these different sources of evidence track differences in which kinds of conclusions the evidence supports. But there is no useful *general* distinction between difference-making and mechanistic evidence-gathering methods, and attempting to construct such a general distinction merely blurs across more useful cross-cutting distinctions at a less abstract level.

It is important that defenders of mechanisms in causal inference recognise this, since it means they are defending the usefulness of a distinction in objects of evidence, without offering any extra evidence-gathering methods. In section 2.2 I will begin examining why evidence *of* mechanisms is so useful for causal inference, and develop that theme throughout the paper.

2.1 Types of evidence 1: Evidence-gathering *methods*

If we are going to categorise types of evidence, the best strategy is to categorise types according to what we use evidence for. I suggest that the best categories for types of evidence are ones that track properties relevant to the *conclusions supported* by each type of evidence. Now, there are several distinctions among types of evidence-gathering methods that matter to the types of conclusion—specifically, the kinds of *causal* conclusion—that type of evidence can support.

At the most specific level, we distinguish types of evidence by differences in the tools or techniques used to gather that evidence. For example, we distinguish: x-ray crystallography evidence, electron microscope evidence, biopsy evidence or blood test evidence, and of course many others. These evidence gathering methods support different kinds of conclusions, specific to what is known about each tool, since any standard tool or technique has its established strengths, and also its limitations. For example, different kinds of spectroscopy give different kinds of information. Infra-red absorption spectroscopy and Raman spectroscopy both give information about vibrational modes of molecules, but are complementary because the modes of vibration that show up by infra-red don't show up by Raman, and vice-versa. UV-visible spectroscopy is complementary again because infra-red and Raman spectroscopy give you wavelengths that correspond with quantum changes in the *vibrational* energy of the molecules, whereas UV-visible spectroscopy gives you absorption bands that correspond with quantum changes in the *electronic* energy of the molecules. These techniques can only be used effectively with an understanding of their limitations. So this category does track differences in conclusions supported by evidence.

Distinctions among evidence-gathering methods at a higher level of abstraction might also be useful for discriminating kinds of conclusions supported. Consider for example:

1. quantitative vs qualitative

2. generic vs single-case

3. evidence that requires large numbers of repeated trials vs evidence that merely requires confirmation

Trials which make no attempt to measure the extent of the effect of a cause might provide some evidence towards supporting a qualitative claim: drug X has some effect on disease Y. But it is reasonable to suppose you need quantitative evidence to support a quantitative conclusion: drug X cures a certain percent of cases of disease Y. In a similar way, single-case evidence, such as a personal manipulation of an experimental set-up, might provide pretty good evidence for a single-case causal claim: my pulling the lever on the toilet caused the flow of water, right here, right now. But repeated trials are generally required to support a generic causal claim, particularly in the biomedical sciences.

The third distinction is interesting. Normally, biomedical trials involve extensive repetition: RCTs must recruit enough participants for the results to be taken seriously. But some experimental work is an exception. The clearest case is breakthrough work in biochemisty. Here, progress often jumps forward due to brand new technology that allows scientists to see biochemical structure that could not be examined before. This is the case for Franklin's x-ray crystallography photographs of DNA, or the use of electron micrographs to see replication forks in DNA, for example. Many unsuccessful trials finally lead to success—a small number of clean images. In such cases results are not extensively repeated—although of course they may require confirmation by a separate research group to check for fraud or bias. In such cases the relatively direct view of structure allows a relatively straightforward choice between competing theories, since it is reasonable to assume that different pieces of DNA are structured in the same way. (Although for discussion of complexities involved in *getting* such clean images in the first place, see Bechtel (2006).)

So it seems that the distinctions above will also track common properties relevant to the conclusions supported by the types of evidence. It is tempting, but far too quick, to move to a yet higher level of abstraction and think that mechanistic evidence-gathering methods are methods that are qualitative, single-case, and let you see pretty straightforwardly that something must be the case; which can be opposed to difference-making methods which are quantitative, generic, and require large numbers of repeated cases.

This move does nothing more than lump together three useful distinctions in types of evidence-gathering methods, and it is potentially very misleading. First, the lumping makes it appear that there are only two categories—mechanistic, and difference-making—combining these three distinctions, when in fact there are many possible combinations. One might well have single-case quantitative evidence, for example. This and further possible combinations are obscured by creating two more abstract categories. Secondly, this lumping might be interpreted as suggesting that evidence *of* a mechanism is always got by evidence gathering *methods*: that are qualitative, single-case, and let you see pretty straightforwardly that something must be the case. This is not true. Nobody would deny that evidence *of* a mechanism can be got by breakthrough technology such as Franklin's photographs. But evidence of a mechanism can also be

got by repeated experimental interventions, as was done in Crick and Brenner's 1961 work using chemical mutagens to crack the genetic code. See Bell (2008) for discussion. We will see further in section 2.2 that the *same* experimental methods are frequently used to gain evidence of mechanism and evidence of difference-making. This means an attempt to separate two divergent categories of mechanistic and difference-making evidence-gathering methods is likely to lead to confusion.

So while there are distinctions in evidence-gathering methods that usefully track something important about the kinds of conclusions the evidence supports, there is *no useful general distinction between mechanistic and difference-making evidence-gathering methods*. Note that the description of categories here is not meant to be exhaustive: there are further useful categories. The best-known categorisation of evidence-gathering methods is by type of study: double-blind randomized controlled trial, controlled trial, observational study, longitudional case series and so on (Higgins JPT, 2009). There are many distinctions here, which should be combined only with care.

2.2 Types of evidence 2: The *object* of evidence

Evidence of difference-making matters. I don't know of anyone who denies that. I will argue here that evidence of mechanism *also* matters. These two types of evidence are complementary: evidence of mechanisms and evidence of difference-making each address different problems for causal inference. This is an idea that has been gaining ground (Russo and Williamson (2007), IARC (2006), Steel (2008)), but I will characterise the complementarity differently.

Evidence of difference-making is just evidence that the effect does indeed vary with the postulated cause. The most familiar cases are correlations found in an RCT or observational study, but evidence of difference-making also includes more complex probability distributions such as those associated with Bayes' nets; simpler relations, which might be single-case, such as a counterfactual relationship (if I hadn't pulled the lever, the toilet wouldn't have flushed); or the more sophisticated invariant relationships Woodward (2003) advocates. Evidence of such relations can be got experimentally in a wide variety of different familiar ways, including at least: clinical trials and observational studies; observing the results of a few cases of simple or experimental manipulations, or of repeated simple or experimental manipulations; and perhaps in some cases from observing the results of either physical or virtual simulations of a system. These kinds of experimental work all give you evidence of a relationship between the postulated cause, and the effect of interest.

The general problem for inferring the existence of a causal relation from evidence of difference-making is well known. When a correlation is found, there are three possible explanations: the existence of a causal relation, accident, or confounding. Even in the most careful experiments, such as in fully randomized double blind clinical trials using large samples, it is difficult to be *absolutely certain* that the observed result is caused by the experimental intervention, and not by confounding: by something that is correlated with the experimental

intervention (Higgins JPT, 2009, section 9.6.5.6). This problem is of course also present to a much greater degree in other kinds of studies, such as observational studies, which are ubiquitous in the health sciences. Without some reason to believe that confounding is absent, the inference to 'C causes E' cannot be made directly from evidence of difference-making between C and E.

Evidence of mechanism is just evidence of the existence of a mechanism or mechanisms in the domain of inquiry in question. The most obvious examples in the biomedical sciences are the various biochemical pathways. Following Illari and Williamson's development of the debate in the philosophical literature on mechanisms, I will take the following definition: 'A mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon.' (Illari and Williamson, 2011, p1.) On this view, evidence of a mechanism is evidence of the entities or activities that make up mechanisms, or the organization of those entities and activities by which they produce the phenomenon the mechanism is known for. Evidence of the mechanism of protein sythesis, for example, would be got by finding evidence of entities involved, such as DNA, RNA of the various kinds, ribosomes, and so on, or the activities in which they engage such as transcription, or regulation and so on.

Just as for evidence of difference-making, evidence of a mechanism—entities, activities, and organization—can be got empirically in many ways, including: direct observation, simple manipulation, repeated experimental manipulation, simulation. For further discussion see Darden (2006); Craver (2007); Bechtel (2008), or Bell (2008) for discussion of Watson and Crick's use of chemical mutagens to crack the DNA code. The key point is there is no principled distinction between the *kinds* of empirical work by which we get evidence of mechanisms, and evidence of difference-making—although in practice for any particular case in the health sciences these different items of evidence are usually got from different studies. In spite of there being no difference in principle in the kind of empirical work done, the difference in the kind of thing we get evidence of—the object of evidence—matters.

In this section I will focus on the role of evidence of a mechanism *linking* the postulated cause with the effect of interest, although in the next section, section 3, I will also examine how general mechanistic evidence about the domain can help in causal inference concerning C and E.

If you have evidence of a mechanism linking C to E—evidence of entities, activities and their organization in the right place at the right time—you have some reason to believe that there is a causal relation between C and E. This is because you have traced the mechanism linking C to E. You have traced the causal process, in Steel's terms (Steel, 2004, 2008). Take protein sythesis again. Once you have found the entities and understood what they do—their activities—and how they are organized relative to each other and the cell, you can trace the path from DNA through replication, to transcription and the various forms of RNA and their activities, to the ribosome and the creation of the amino acid chain. Since you have traced the causal process that begins with the DNA and ends with the required amino acid chain, you can be confident

that one can begin with DNA and end up with an amino acid chain.

But you *cannot* yet conclude that C causes E. The general problem for inferring 'C causes E' from evidence of a mechanism linking C and E is the problem of masking. You have found one link from C to E, but you do not know what other links there may be. Suppose you have found a link by which C increases E. There may still be another link, a route by which C *reduces* E. Steel gives the relationship between exercise and weight loss as an example (Steel, 2008, p68). Increased exercise leads to more calories being burned, via well-known mechanisms. But it also leads to increased appetite. Until we investigate further, we do not know the overall effect of exercise on fatness or thinness, and we cannot claim either that exercise makes you fat, or makes you thin. The operation of one mechanism might *mask* or hide the operation of the other.

It might be objected that C still causes E if there is a linking mechanism, whether or not that mechanism is masked. There is still a 'component effect', even if there is no 'net effect', in Hitchcock's terms (Hitchcock, 2001).[1] To use Hitchcock's example, taking birth control pills directly causes thrombosis, and also indirectly prevents thrombosis by preventing pregnancy, which itself causes thrombosis. In such a case, the state of pregnancy is sufficiently different from the state of non-pregnancy to make a reasonably clear distinction between the two mechanisms and maintain the separate causal claims. The problem is that there are many cases in the biomedical sciences where this is not clear, due to the complexity and integration of regulation and control mechanisms which characteristically act on each other within both cell and body. Multiple causal pathways between variables of interest may well be the norm. This means there may be many cases where we can trace a mechanism between variables, but still have no idea whether the first variable increases, decreases, leaves untouched, maintains in homeostasis, and so on, ... the other. Causal relationships certainly exist in the local domain, but in such circumstances to assert a causal relationship *between those variables* would be very misleading. Complementary difference-making evidence is required.

So even though the same experimental techniques are used to get evidence of difference-making and evidence of mechanism, the difference matters. This is because evidence of a difference-making relation between C and E, and evidence of a mechanism between C and E, are not just independent 'pluses' in favour of the causal claim 'C causes E'. This is because evidence of difference-making and evidence of mechanism integrates: *each addresses the major weakness of the other evidence.* Together they are much better evidence for the existence of a causal relation than evidence either of difference-making, or of mechanism, can be on its own.

Finding a mechanism is one good way of increasing confidence that any relationship between C and E is *not due to confounding* (or, indeed, due to chance). Various techniques, particularly randomizing and blinding, are also very effective at increasing this confidence. I certainly advocate their rigorous

[1] I owe thanks to an anonymous referee for pressing me on this point.

use, wherever possible. But evidence of a linking mechanism is a *distinct* source of such confidence currently neglected by, for example Higgins JPT (2009). If you can trace a link all the way from C to E, it is considerably less likely that any relation between C and E is merely due to a correlation between C and the real cause of E, and considerably less likely to be due merely to chance. I will address other ways in which mechanistic evidence can help with confounding in section 3. On the other hand, finding a difference-making relation is the best way to deal with the problem of *masking*. Finding an overall difference-making relation between C and E is a good way of increasing your confidence that the link you have traced between C and E is not masked by other undiscovered routes between C and E.

So in favour of the claim 'C causes E' we have:

1. Evidence of a difference-making relation between C and E:

   Problem is confounding

   Advantage is eluding masking

2. Evidence of a mechanism linking C and E:

   Problem is masking

   Advantage is avoiding confounding (and chance)

So the best evidence you can have of a causal claim 'C causes E' is good quality evidence of the strength of a difference-making relation between C and E, such as that got from a well-conducted RCT; *plus* evidence of a mechanism linking C and E that seems to accord with the strength of the difference-making relation detected between C and E, such as that got from a well understood biochemical pathway. The mechanistic evidence is not negligible here, and when the best available evidence of difference-making is from an observational study, it is even more important.

It is useful to expand a little more on the claim that there is no principled difference in the kinds of experimental work that we use to get evidence of difference-making, and evidence of mechanisms. It might be objected that surely we don't conduct a single trial, take the results and call them evidence of difference-making, then take the same results and call them evidence of mechanism, then add the two together and claim to have better evidence for a causal claim than we ever thought we could get out of a single trial.

This is not what I intend. In most cases, particularly in the biomedical sciences, causal conclusions are arrived at as a result of many trials. The EBM hierarchy (or hierarchies) and the IARC are both absolutely right to put conclusions of combined work ahead of the results of a single trial. The framework I have provided here is intended partly as a framework for understanding more formally the integration of work from different trials—how to approach the results and come to an overall conclusion on their basis. Particularly, we can see how work from classic studies aimed at establishing difference-making, such as RCTs, integrates with work clearly aimed at establishing knowledge of mechanisms, such as in vitro studies or in vivo studies in animals aimed at exploring

biochemical pathways. Each has a clear function in building up an overall understanding of the causal picture, and this framework allows it to be clearly articulated.

It is *in principle possible* that a single experiment could establish a causal claim. Psychological work on small children investigating simple mechanical toys shows that they come to an accurate picture of the causal structure very quickly, particularly if they are allowed to manipulate the mechanism themselves. (See Gopnik and Schulz (2007) for debate.) Using the framework, we can say that they acquire both evidence of difference-making (by seeing, say, that pulling the lever flushes the toilet) and evidence of mechanism (say, by observing the cistern, and the attachment of the lever to the valve) almost simultaneously. This kind of easily accessible causal knowledge may be an important feature of daily life, particularly in our interactions with simple artefacts, but it will be extremely rare in the far more complex domains of the biomedical sciences. Instead, much experimental work gives us a rich array of different evidence, of mechanisms and of difference-making, especially as evidence accumulates over a variety of studies in a field.

So Russo and Williamson are right to draw attention to the importance of mechanisms in causal inference. This is due to the different roles evidence of mechanisms and of difference-making play in causal inference.

3 Second Distinction: Evidence of *what* mechanism(s)?

I have offered good reason to be interested in the use of evidence of mechanisms in causal inference. In the rest of this paper, I will unpack that idea further, examining in this section the question of evidence of *what* mechanism or mechanisms might be required, and in section 4 the question of *how much* evidence of a mechanism or mechanisms might be required.

This brings me to my second distinction. There are two different categories of mechanisms, evidence of which gives you support for a causal inference 'C causes E' in different ways.

Second Distinction

1. Evidence of a mechanism *linking* C and E.
2. Evidence of other mechanisms, usually in the domain where C and E are found.

I have explained above how evidence of a mechanism *linking* C and E can be important for the inference to 'C causes E'. Note that such evidence is very specific. It is evidence of activities, entities and their organization *actually linking the particular postulated cause and effect*. It is this that allows a link to be traced between cause and effect. But evidence of *other* mechanisms might bear on the question of whether C can cause E. For example, suppose

we are wondering whether a particular bacteria causes a disease in human beings. Evidence of mechanisms actually active in human beings linking similar bacteria to similar kinds of diseases—evidence of *analogous* mechanisms in human beings—will have a bearing. For example, the discovery of the mechanism by which *Clostridium tetani* causes tetanus in human beings was important in understanding analogous mechanisms of disease causation. Infection with *Clostridium tetani* is highly local, often to a particular wound, but its effects are muscle spasms in distant parts of the body. It was discovered that *Clostridium tetani* releases a very toxic protein, 'tetanus toxin' which escapes from the cell to travel over the body, causing the muscle spasms. The fact that bacteria can cause disease by sitting still and releasing a toxin, rather than by invading the body, was important in discovering the mechanism for analogous diseases, such as botulism. *Clostridium botulinim* remains in the gastrointestinal tract, but also releases a toxic protein, 'botulinum toxin', which paralyses muscles. Evidence for the existence of a mechanism linking the bacteria to the disease in chimpanzees, and in apes, can also bear on our inference, as in this case. Note that in some cases this might count as evidence for the existence of a linking mechanism in human beings, when we have reason to believe that the human mechanism and the mechanism in our close relatives is the same. But even when it is not the same mechanism, such evidence still bears on the issue. So these are examples of cases where relevant evidence includes evidence of mechanisms in the domain where the cause and effect of interest are also found, but are not themselves evidence of mechanisms actually linking C and E.

Note that evidence of one kind of mechanism(s) can be had without the other. We may know a reasonable amount about the mechanisms of a domain, but know very little about whether there is a mechanism linking two particular variables in that domain. It is conceivable that you could have some evidence of a mechanism linking two variables in a domain, but otherwise know very little about the domain. Note that evidence of mechanisms in the domain can also be negative: evidence of the *absence* of mechanisms which have been postulated, searched for, but found to be not there.

In this section I examine how evidence of mechanisms that do *not* link C and E might nevertheless bear on the inference to C causes E. I will argue that there may be cases where it is *possible* to establish that C causes E without evidence concerning the mechanism linking C and E. So I deny *this interpretation* of the RWT. (Although section 4 will further clarify this claim in the light of the third distinction.) Nevertheless, I will argue, it is plausible that you cannot infer C causes E without *some* mechanistic evidence, even if this is only of other mechanisms. So I support this, far weaker, interpretation of the RWT.

To reiterate, the problem with even the best quality evidence of a difference-making relation between C and E is that there are three possible explanations for it: causal relation, chance, or confounding. I have argued above that evidence of a linking mechanism helps increase confidence that the correlation found is not due to confounding, or unlucky chance. But evidence of other mechanisms can also help with this. Evidence of other mechanisms in the domain tells you about *other* links that exist in the domain, while work aimed at discovering

these other mechanisms also tells you something about links that *don't* exist in the domain.

In this way, evidence of *other* mechanisms—mechanisms other than those linking C and E—still helps us delimit the range of possible causal links in a domain. The better understood the domain, the more substantial this knowledge will be. In a well-understood domain, you have a much better idea of the variables that might be confounders— variables that might themselves cause the effect of interest, while also being correlated with the postulated cause. And the more mechanisms that have been discovered—the more thoroughly you have understood the domain—the more confident you can be that there are no more *undiscovered* mechanisms in that domain. Alternatively, of course, such thorough understanding might be what alerts you to the fact that you are dealing with a complex system, with causal relations highly ubiquitous and subject to altering with small contextual changes, so that the problem of confounders is not likely to be successfully dealt with at all. This tells you that you must be cautious in making causal claims.

This kind of knowledge is exactly what is needed to analyze the method and results of an RCT, or an observational study. It is what allows you to compare treatment and control groups and be as confident as possible that they do in fact resemble each other with respect to all *relevant* factors. This might in some cases allow you to be confident enough to come to a causal conclusion on the grounds of the good difference-making evidence you have, plus the solid evidence concerning general mechanisms in the domain which allows you to be confident that confounders are *absent*. This kind of evidence is characteristically brought to bear in designing, and in analyzing the results of studies, including RCTs, and so it is a mistake to think that a well-conducted RCT supports a causal claim without any reliance on mechanistic evidence whatsoever. Mechanistic evidence is implicitly used to secure the conclusion that the RCT has been successfully conducted.

An objection to the RWT in the literature comes from Alex Broadbent (Broadbent, 2011, p62). It is that the requirement for evidence of mechanism will bias scientific research towards existing knowledge. This is because we will look for mechanisms like the ones we have already found. I think that this is a legitimate—and indeed often operative—requirement on scientific method. If a new finding is inconsistent with other things we know about the domain, we should be suspicious. Indeed, it is often recognised that background knowledge plays an important part in our decisions about what studies to perform, and what conclusions to draw. But this is often stated very informally. The framework presented here provides a promising avenue for stating this requirement more precisely, hopefully leading to fruitful work—both empirical and theoretical—on when existing knowledge is a fruitful constraint, and when there are indicators that we are about to discover something fresh, and should abandon the constraint. I do not advocate merely ignoring good difference-making evidence, just integrating it with good evidence of mechanism.

In conclusion, without some reason to be confident that there are no confounders, evidence of a difference-making relation between C and E does not

establish a causal claim. Evidence of mechanisms—both linking and other mechanisms—can be a good reason for being confident that there are no confounders. On the other hand, finding a mechanism linking C and E tells you nothing about other links between C and E, and so not enough about the overall strength of any link between C and E. Evidence of difference-making is required to address this problem. So Russo and Williamson's claim that causal inferences require evidence both of difference-making, and of mechanisms, might refer *either* to a mechanism linking C and E, *or* to other mechanisms in the domain. This second possibility is not normally considered, and the claim that establishing causal claims requires evidence of some mechanisms in the domain is a much weaker claim than Russo and Williamson are often interpreted as making. I will now go on to develop my third distinction, before returning to the Russo-Williamson thesis for the final time.

4 Third distinction: How *much* evidence?

The third distinction is in how much evidence of a mechanism might be required to establish a causal conclusion. These different amounts of evidence might apply to our evidence of any target mechanism, linking or otherwise. Some of the discussion of the RWT does not explicitly clarify this question, leading to confusion. There are at least four relevantly different cases:

Third Distinction

1. Evidence of *what the mechanism is* in detail.
2. Evidence *that there is* a mechanism of the postulated kind.
3. Postulated mechanism, based on evidence of analogous mechanisms.
4. Evidence that there is *no* mechanism.

Clearly these start at the most demanding, and so the best evidence. But this is not a simple continuum of decreasing amounts of evidence of a mechanism. There are two interesting changes in kind of evidence: between cases two and three, and again between cases three and four.

I will focus on the case where we are investigating a linking mechanism between C and E, so the linking mechanism is the target mechanism. Suppose again that we are interested in whether a particular bacterium causes a particular disease in human beings. We are in the first position when we have pretty good detailed evidence about exactly what kinds of entities and activities link *bacteria* and *disease*, in human beings, and how their organization produces the difference-making relation between *bacteria* and *disease* that we observe—including if at all possible the strength of that difference-making relation. In the tetanus case, the most important evidence was isolating the toxic protein, 'tetanus toxin' and establishing that it escapes from the cell. The second case is where evidence is less detailed. We may be unsure about some of the entities

and activities, but have enough evidence of enough of them to be pretty sure that there is a mechanism of the postulated kind. In the botulism case, with an analogous mechanism of disease causation in mind, searching for another toxic protein was a natural step, and isolating 'botulinum toxin', discovering that it could be in food before ingestion, or produced within the gastrointestinal tract, and then escape into the bloodstream to have remote effects, were all important steps. Clearly both of these involve positive evidence of the mechanism actually sought. Naturally it is desirable that this mechanism cohere with any other mechanisms in the same domain that we have evidence for, but the primary source of evidence is of the mechanism actually linking C and E.

The third case isn't just yet fuzzier evidence about the mechanism linking *bacteria* and *disease*. It may involve some basic evidence of entities and activities linking *bacteria* and *disease*. But the primary source of evidence for the mechanism of interest shifts—the primary source is now evidence of *other mechanisms*. So we clearly have evidence of many relevant metabolic mechanisms in human beings, and we might also have evidence of the action of bacteria or toxic protein in animals, as well as in vitro evidence of the production of the toxin by the bacteria. In the botulism case, the bacterium was identified, and the tetanus case provided evidence of an analogous mechanism of disease causation in humans. Both tetanus toxin and botulinum toxin can be produced artificially in culture in vitro. When the bacteria are separated out, and the remaining liquid injected into mice, the mice develop either tetanus or botulism. This all amounts to considerably more evidence than the simple absence of positive reason to believe there is no mechanism. The important question is whether a postulated mechanism of interest in human beings can be constructed that accords reasonably well with our evidence of other mechanisms in the domain. This requires coherence with the kinds of activities and entities and their organization that we generally find in the domain; and also with what we know about likely and unlikely causal links in the domain. But if we have this, we may be reasonably sure *some* such mechanism exists. This case is important because, although you don't have positive evidence for the exact linking mechanism in human beings, in certain cases you might be pretty certain that such a mechanism exists. Of course, you might *not* have such evidence, in which case you cannot make a causal inference, even with some difference-making evidence. But case three correctly identifies what kind of evidence is relevant to the judgment about whether some such mechanism exists or not.

The fourth case is where you can't see how there *possibly could be* a mechanism linking C and E. This comes entirely from knowledge of other mechanisms in the domain, or of course from any other domain in science that bears on the issue. We may know that anything linking C and E would be so unlike the other mechanisms we find in the domain, that it is not plausible. So we cannot even think of a plausible postulate for a mechanism linking C and E based on reasoning from analogy from other mechanisms in the domain. For example, this is an important source of evidence bearing on whether homeopathic remedies cause recovery. We can currently envisage no plausible mechanism by which such remedies would work, that fits with experimental evidence, or what we

know about the physical chemistry of water.

For cases one and two the primary source of evidence is of the target mechanism itself, the mechanism linking *bacteria* and *disease*, while for cases three and four the primary source of evidence of the target mechanism comes from evidence of other mechanisms. This means that for both the third and fourth cases, the evidence available can be considerably stronger or weaker, depending on how much is known about the domain. Suppose you have a postulated mechanism, and a little bit of fuzzy and ambiguous evidence of some activities or entities analogous to those postulated linking *bacteria* and *disease*. What do you know? This depends on how much is known about the domain.

If you don't know very much about the domain, this kind of evidence isn't going to say much. It is certainly not going to rule out a causal relation between *bacteria* and *disease*. But on the other hand, you lack *either* a solid mechanism linking *bacteria* and *disease*, *or* enough background evidence of mechanisms in the domain to rule out other possible links—other possible confounders— or the resulting correlations being due to chance, with any confidence. So you cannot confidently move to a causal claim either. The more you know about the domain, the better this kind of evidence is. If you know a great deal about the domain, you may still be able to rule out possible confounders and chance with reasonable confidence, and think you are well on the way to a solid causal claim on the basis of good difference-making evidence.

Indeed, in a very well known domain, no evidence of a mechanism linking C and E might count as strong evidence that there is *no* mechanism linking C and E. This is simply because, in a thoroughly investigated domain, if there were such a mechanism, we would expect to have found *some* evidence of it by now. For example, the links and possible links in a simple mechanical child's toy can be so easy to exhaust, and our understanding of simple mechanical interactions so good, that if a parent carefully examining it failed to notice a link between A and B, that is excellent evidence that there is no such link. This might also happen in a well-known domain when we had one or more postulated mechanisms, but exhaustive searches have failed to find any of the expected entities or activities. We are sure that this mechanism or these mechanisms do not exist. In a much less well-known domain, such as complex systems, the same experimental results might leave us still unsure whether the postulated mechanisms we were looking for exist or not.

Although cases three and four both involve evidence of other mechanisms in the domain bearing on whether there is a mechanism linking C and E, case four is different from the other cases because it allows the possibility of *positive* evidence of *absence*. This possibility might be doubted, but some clear cases can be found. For example, in many cases physics rules out the possibility of any mechanism. There are some fundamental rules we think we know from physics, that have withstood testing for a long time, such as the maximum speed of light. Since causal influence of any kind, such as light, electromagnetic radiation, sound waves, or any moving body, cannot travel faster than the speed of light, *anything* that is a candidate cause of E has to be in the backwards light-cone of E, since the backwards light-cone delimits the range of space-time

from which something could have travelled to E at the speed of light. This is a powerful limit on possible causal relations coming to all disciplines from physics. This kind of positive evidence of absence can be powerful enough to rule out a causal relationship entirely. However convincing any putative difference-making evidence between E and a candidate cause that is not in the backwards light-cone of E, it must be due to error or coincidence. (We might be able to rule out confounding in such a case, since a possible confounder might also fail to be in the backwards light cone of E.)

This might sound like an otiose case. But actually it is common practice to use results from different disciplines to frame our ideas about possible causal relations, for all that some of the most fundamental constraints are usually thought too obvious to state. The constraint from physics most familiar in the biomedical sciences is probably energy constraints. If a postulated reaction in a biochemical pathway requires more energy than is available, it is ruled out. It must be abandoned, or the hidden source of energy found. But it is physics that tells us that energy is conserved: it doesn't come from nothing, or vanish into nowhere. These constraints, when we find them, are very important. The framework developed here allows us to articulate clearly how they impact on biomedical research: they are best characterised as evidence that there could be no possible mechanism, or, in the energy case, no possible mechanism of the postulated kind. At a deep level, such evidence frames our whole approach to finding causes.

So we have four different cases of evidence of mechanisms bearing on causal inference, differing in the amount and source of evidence. All of these kinds of evidence of mechanisms are helpful in causal inference, but they are not all required.

5 The Disambiguated Russo-Williamson Thesis

With these clarifications in place, we can turn once more to the RWT. I have argued in section 2 that 'mechanistic evidence' should be interpreted as 'evidence of a mechanism', and in section 3 that the relevant mechanism we seek evidence of might be a mechanism linking C and E, or other mechanisms in the domain. I have made it clear that if the RWT is interpreted as holding that a distinctive mechanistic evidence gathering method is required for causal inference, then I deny it, and I have indicated that Russo and Williamson also deny it. In section 3, I said, as a first approximation, that it is plausible that we always need evidence of mechanisms in the domain or some other domain to establish a causal claim, but that it was possible that we could establish 'C causes E' without evidence of a *linking* mechanism. I am now in a position to clarify further this much weaker sense in which mechanistic evidence *is* required for causal inference.

First, it is worth noting that 'establishing' a causal claim does not imply its truth. The causal claims that concern me are warranted scientific claims, and scientific claims are *permanently* open to revision by science. This is the

nature of science. This is a source of confusion over the original RWT. Top quality difference-making evidence, plus detailed evidence for the particular mechanism linking C and E, that accords with the strength of the difference-making relationship, establishes that C causes E, but this does not imply that this claim could not be revised in the light of later work. As Bradford Hill writes:

> All scientific work is incomplete—whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time. (Hill, 1965, p300.)

Even moving to the correct understanding of 'mechanistic evidence' as evidence *of* mechanisms, I have denied two stronger senses of the RWT. It is not plausible that detailed evidence of what the mechanism is linking C and E—case one—is required to establish a causal claim. Of course, when you have it, such evidence is very helpful, and may further secure your causal claim. It is plausible that you *frequently* need evidence that there is some mechanism or other linking C and E—case two—to establish a causal claim. This needs slight care, because of course what is desired is evidence *over and above* the simple fact of a difference-making relationship. The existence of a difference-making relationship alone is prima facie evidence of the existence of a mechanism. This is why such evidence sends us scurrying off looking for a mechanism. But of course *this* evidence of a mechanism alone doesn't give us better support for the causal claim than the difference-making relationship itself. *Independent* evidence of the mechanism is needed. Such evidence is not *always required* for causal inference if excellent difference-making evidence is available, but it is very helpful. The amount of evidence of a mechanism described in both cases one and two can be decisive when available difference-making evidence is not top-quality.

What remains is two much weaker senses in which I hold that the RWT is true. Cases three and four describe amounts of evidence of some mechanisms in the domain, but not a linking mechanism, that bear on whether C causes E:

1. Case three: Evidence that supports a postulated linking mechanism, based at least on evidence of other mechanisms in the domain, allowing you to rule out some possible confounders of the difference-making relationship between C and E.

2. Case four: *No* positive evidence of the *absence* of a mechanism linking C and E.

Note that neither of these cases involve detailed evidence of a linking mechanism. I will take case four first. Evidence that there is no mechanism, as I have said, can come in the strong form above where fundamental sciences rule out possible mechanisms entirely, and in weaker forms particular to the biomedical

sciences, where possible mechanisms look very implausible, based on evidence concerning other mechanisms in the domain. The strength of such evidence depends on how much is known about the domain. It will also demand a certain amount of consensus. A single scientist insisting that there could be no such mechanism twenty years after the rest of the field believes there is such a mechanism does not count as evidence that there is no mechanism. Case three is the weakest form of positive evidence. It can be very vague, so long as it allows some judgments about possible confounders that allows you to have some confidence in the difference-making evidence available.

It might be objected that these forms of mechanistic evidence are trivial. It is true that we almost always have such evidence in biomedical cases—it is often called 'background knowledge'. However, there are reasons for recognising that these are forms of mechanistic evidence. If the question is whether they are difference-making evidence or mechanistic evidence, then they are clearly mechanistic. Further, we may well wish to discriminate amongst different forms of *background* knowledge, not all of which might be mechanistic evidence. These kinds of evidence play the same role of other kinds of mechanistic evidence, and are rightly classified alongside them.

Daniel Steel has objected to what he calls the alleged negative role of mechanisms. He says that using mechanisms in causal inference cannot work, because one can always think of a plausible mechanism Steel (2004). Looking at cases one to four undermines this worry. Steel is concerned that postulated mechanisms amount to no more than what he calls 'how-so stories' about mechanisms. But I have shown that there is much more than this to the story about how mechanisms are important to causal inference. Most importantly, mechanistic evidence is never just a story, but consists of *evidence supporting* a story giving a postulated mechanism. Note that Steel is also underestimating the importance of case four. It is important that we do not have positive evidence that there is no such mechanism, and I envisage case three as including *not* being in case four.

Steel is a philosopher of social science, and is particularly concerned about how-so stories about mechanisms in the social sciences. He seems to be right that this is a bigger problem in the social sciences than the biomedical sciences. It is harder to rule out mechanisms as implausible there based on what else is known about mechanisms in the domain. The interesting question is whether this is merely because the mechanisms in the social sciences are less well understood than, say, biochemical pathways. Perhaps when we understand social mechanisms better, the problem will be less; or perhaps it is due to the peculiarly contextual nature of social mechanisms, and so is an ineliminable problem of the domain. Only time will tell.

In these two much weaker senses, evidence of mechanisms is needed to establish causal claims.

6 Conclusion

I have thoroughly examined the ways in which mechanistic evidence can help with causal inference, when combined with difference-making evidence. I have introduced three distinctions: mechanistic evidence as evidence-gathering method versus object of evidence; evidence of a mechanism linking C and E versus evidence of other mechanisms in the domain; and four different cases for how much evidence of mechanisms is available.

I have argued that mechanistic evidence is not useful as an evidence-gathering category. I have argued that even when we understand the distinction as between *objects* of evidence, evidence of mechanisms is helpful for causal inference, since it performs a complementary job to evidence of difference-making. Evidence of mechanisms helps us avoid confounding, while being open to the problem of masking; evidence of difference-making helps us avoid masking, while being open to the problem of confounding. I have examined how evidence of a linking mechanism between C and E can help with causal inference, as can evidence of other mechanisms in the domain by indicating possible and impossible causal links. I have examined four relevantly different *amounts* of evidence one might have of mechanisms, and illustrated how even the weaker can still be useful for causal inference.

I have illustrated how my framework using the three distinctions allows greater precision in describing what evidence for a causal claim is available, and what might still be needed. I hope that this is a fruitful place to start to ask more precise questions about the place of mechanisms in causal inference. It is worth reiterating finally that mechanisms have been advertised in the literature as able to help with other jobs such as explanation, and estimating the stability of causal relations, and I have not examined these. In particular, I have not here considered the tricky issue of external validity, which I intend to turn to in further work.

Acknowledgements

I am grateful to the Leverhulme Trust for supporting this research. I am also indebted to colleagues at Kent and in the Causality in the Sciences network for discussion of many of these issues. Particular thanks are due to Lorenzo Casini, Brendan Clarke, Donald Gillies, Federica Russo, Attilia Ruzzene and Jon Williamson. Remaining errors are, of course, my own.

References

Bechtel, W. (2006). *Discovering Cell Mechanisms: the Creation of Modern Cell Biology*. CUP, Cambridge.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge, Oxford.

Bell, G. (2008). *Selection: The Mechanism of Evolution, (2nd Edition)*. OUP.

Broadbent, A. (2011). Inferring causation in epidemiology: mechanisms, black boxes, and contrasts. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 45–69. OUP.

Craver, C. (2007). *Explaining the Brain*. Clarendon Press, Oxford.

Darden, L. (2006). *Reasoning in Biological Discoveries*. Cambridge University Press, Cambridge.

Gillies, D. (2011). The Russo-Williamson Thesis and the question of whether smoking causes heart disease. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 110–125. OUP.

Gopnik, A. and Schulz, L., editors (2007). *Causal learning: psychology, philosophy, and computation*. Oxford University Press, New York.

Higgins JPT, G. S., editor (2009). *Cochrane Handbook for Systematic Reviews of Interventions*. Available from www.cochrane-handbook.org. The Cochrane Collaboration, version 5.0.2 [updated september 2009] edition.

Hill, B. (1965). The environment of disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300.

Hitchcock, C. (2001). A tale of two effects. *The Philosophical Review*, 110(3):361–396.

Howick, J. (2011). Exposing the vanities - and a qualified defence - of mechanistic reasoning in clinical decision-making. Forthcoming.

IARC (2006). Preamble to the IARC monographs. Technical report, IARC (International Agency for Research on Cancer), http://monographs.iarc.fr/ENG/Preamble/index.php.

Illari, P. M. and Williamson, J. (2011). What is a mechanism?: Thinking about mechanisms across the sciences. Under review.

Kincaid, H. (2011). Causal modeling, mechanism, and probability in epidemiology. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 70–90. OUP.

Leuridan, B. and Weber, E. (2011). The IARC and mechanistic evidence. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 91–109. OUP.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Russo, F. and Williamson, J. (2011). Generic versus single-case causality: the case of autopsy. *European Journal for Philosophy of Science*, forthcoming.

Steel, D. (2004). Social mechanisms and causal inference. *Philosophy of the Social Sciences*, 34:55–78.

Steel, D. (2008). *Across the boundaries. Extrapolation in biology and social science*. Oxford University Press.

Vreese, L. D. (2008). Causal (mis)understanding and the search for scientific explanations: a case study from the history of medicine. *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 39:1424.

Weber, E. (2009). How probabilistic causation can account for the use of mechanistic evidence. *International Studies in the Philosophy of Science*, 23(3):277–295.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. OUP.