University of Kent

Kent Academic Repository

Williamson, Jon (2021) *Evidential Pluralism and Explainable AI*. The Reasoner, 15 (6). pp. 55-56.

Downloaded from
https://kar.kent.ac.uk/92399/ The University of Kent's Academic Repository KAR

The version of record is available from
https://riviste.unimi.it/index.php/thereasoner/issue/view/2403

This document version
Publisher pdf

DOI for this version

Licence for this version
CC BY (Attribution)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

The Reasoner

Volume 15, Number 6
December 2021

thereasoner.org
ISSN 1757-0522

Contents

Editorial **45**

Features: Focus on Evidential Pluralism **45**
- Introducing EP . . . 45
- EP in Medicine . . . 47
- EP and the effectiveness of Treatments in Medicine . 48
- EP and Sports and Exercise Science . . . 49
- EP in the Social Sciences . . . 50
- EP in Social Policy . . . 51
- EP and Econometric Modelling . . . 52
- What counts as evidence for a mechanism? And why? 53
- EP in Cognitive Science . . . 54
- EP and Explainable AI . . . 55

Events **57**

Courses and Programmes **57**

Jobs and Studentships **57**

Editorial

The turn of the new millennium coincided with a mechanistic turn in the philosophy of science. One strand of this new mechanistic philosophy, Evidential Pluralism, provides an account of the epistemology of causality that treats evidence of mechanisms on a par with evidence of correlation when assessing causality. The focus of this issue of *The Reasoner* is on Evidential Pluralism. The following articles introduce the approach and its application to science and practice.

Please consider submitting developments and criticisms of Evidential Pluralism to future issues of *The Reasoner*.

Jon Williamson
Department of Philosophy and Centre for Reasoning
University of Kent

Features: Focus on Evidential Pluralism

Introducing Evidential Pluralism

It is a platitude that correlation does not on its own imply causation. But what else is required to establish causation? Russo and Williamson (2007: Interpreting causality in the health sciences, *ISPS* 21(2), 157–170) argued that one also needs to establish a mechanistic connection between the putative cause and effect.

This is because an observed correlation between A and B might be attributable to one of a large number of different explanations. While it might be that A is a cause of B, there might instead be some other causal connection: perhaps B is a cause of A, or some common cause C is responsible for the correlation. The latter scenario is usually understood as a kind of confounding or selection bias. But there are many other kinds of bias that could account for the correlation—e.g., performance bias and detection bias. Or the correlation could be a statistical artefact: attributable to the size of sample, fishing for correlations, or temporal trends in the data, for example. Finally, the correlation could be due to some non-causal connection between A and B, such as a semantic, constitutive, logical, physical or mathematical relationship between the two variables. What is distinctive about the case in which A is a cause of B is that there is some mechanism of action by which A causes B.



Fig. 1 represents the confirmatory relationships that underpin Evidential Pluralism. The top part of the diagram invokes the above thesis: to establish a causal claim one normally needs to establish both correlation and mechanism, so to assess a causal claim one needs to assess both a correlation claim and a mechanistic claim. The correlation claim is that A and B are probabilistically dependent conditional on potential confounders.

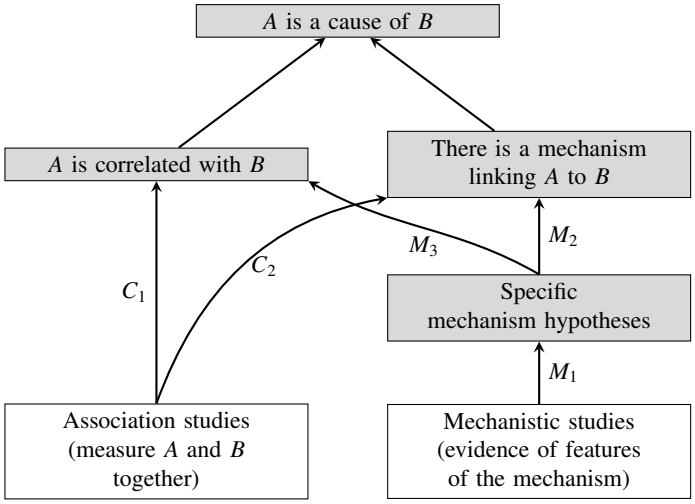


Figure 1: Evidential relationships for assessing a causal claim.

(What counts as a potential confounder may depend on previously confirmed causal claims, confirmed theory, and other evidence.) The mechanism claim is that A and B are connected by a complex of mechanisms that cite A as being responsible for B and that can account for the extent of the correlation. This complex of mechanisms needs to include some mechanism of action from A to B, and any counteracting mechanisms should not wholly cancel out this mechanism of action.

Lower down the diagram, we encounter the evidence required in order to establish correlation and mechanism. The obvious way to test for a correlation is to perform an association study. Here, an association study is an experimental or observational study that makes repeated measurements of A and B, together with potential confounders, in order to determine whether they are associated and, usually, the extent of any correlation. Association studies confirm correlation along confirmation route C_1 in the diagram. Such studies can also indirectly confirm the existence of a mechanism along route C_2. For example, the presence of several large, concordant randomised controlled trials (RCTs) that find a strong association reduces the plausibility of confounding, and thereby confirms the claim that there is some mechanism from A to B that accounts for the correlation.

A more direct way of confirming this mechanistic claim involves hypothesising key features of the mechanism complex: mediating variables, entities, activities and organisational features of the mechanism, for example. If the presence of these features is confirmed, this in turn confirms the existence of the underlying mechanism (confirmation channel M_2). Mechanistic studies are studies that provide evidence of these features. In certain cases, the key features of the mechanism can also make the existence of a correlation more credible (M_3): for example, the features of a parachute mechanism confirm the claim that parachute use will decrease the probability of serious harm when falling from a plane, obviating the need for RCTs.

We thus have:

Evidential Pluralism. In order to establish a causal claim one normally needs to establish the existence of an appropriate conditional correlation and the existence of an appropriate mechanism complex, so when assessing a causal claim one ought to consider relevant association studies and mechanistic studies, where available.

The first part of this thesis, *object pluralism*, specifies two objects of evidence. The second part, *study pluralism*, specifies two kinds of study that constitute the evidence. Thus,

Evidential Pluralism = object pluralism + study pluralism

The need to distinguish object and study pluralism is motivated by Illari (2011: Mechanistic evidence: Disambiguating the Russo-Williamson thesis, *ISPS* 25(2), 139–157). Evidential Pluralism is controversial largely because it goes against the evidential monism enshrined in the orthodoxy of present-day evidence-based medicine and evidence-based policy, which focus on association studies (particularly RCTs) to the exclusion of mechanistic studies (Clarke et al. 2014: Mechanisms and the Evidence Hierarchy, *Topoi* 33(2): 339–360). Parkkinen et al. (2018: Evaluating evidence of mechanisms in medicine: principles and procedures, Springer) developed methods for evaluating mechanistic studies alongside association studies in medicine. Shan and Williamson (2021: Applying Evidential Pluralism to the social sciences, *EJPS*, in press) argue that Evidential Pluralism can be fruitfully applied to the social sciences. The quantitative approach to the social sciences relies almost exclusively on association studies, while the qualitative approach and methods such as realist evaluation, process tracing, contribution analysis and theory of change can be viewed as focusing on mechanistic studies. Evidential Pluralism is an integrative, mixed-methods approach.

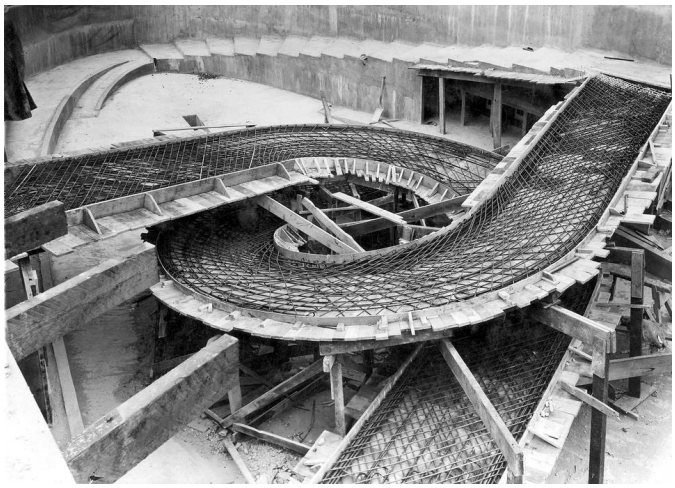


Figure 2: Reinforced concrete in the construction of the penguin pool at London Zoo.

An analogy is sometimes drawn with reinforced concrete (Clarke et al. 2014). Concrete is resistant to compression but fails under tension. Steel, on the other hand, is resistant to tension but buckles under compression. Putting them together in reinforced concrete enables strong structures resistant to both tension and compression. Association studies and mechanistic studies reinforce one another in a similar way. Association studies provide excellent evidence of correlation but are prone to biases and confounding and so provide much weaker evidence of mechanisms. Mechanistic studies provide excellent evidence of mechanisms, but the complexity of mechanisms and the presence of counteracting mechanisms mean that they provide only weak evidence of a net correlation. However, putting them together yields a strong evidential structure that

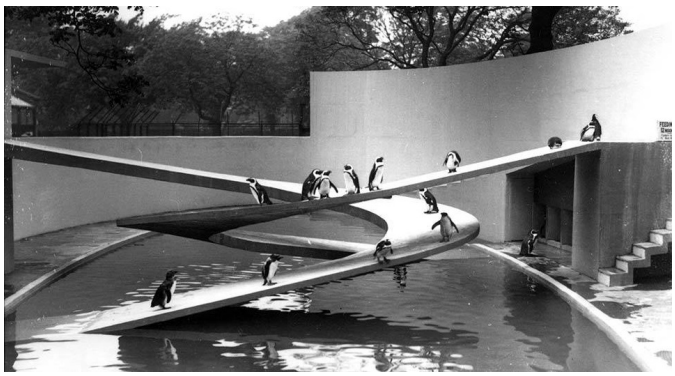


Figure 3: The penguin pool, completed in 1934.

can more easily establish causation.

Jon Williamson
Department of Philosophy and Centre for Reasoning
University of Kent

Evidential Pluralism in Medicine

The above introduction to Evidential Pluralism distinguishes between an *association study* and a *mechanistic study*, and then appeals to this distinction in order to characterize Evidential Pluralism. According to this characterization, establishing a causal claim typically requires evaluating both association studies and mechanistic studies. In this piece, I will give a brief overview of some of the work discussing this Evidential Pluralism in the context of medicine.



This Evidential Pluralism was first put forward in the context of medicine and the health sciences by Federica Russo and Jon Williamson (2007: Interpreting causality in the health sciences, *International Studies in the Philosophy of Science*, 21(2): 157–70). Among other things, Russo and Williamson appealed to a number of case studies to argue that establishing causal claims in the health sciences requires evaluating both association studies and mechanistic studies. For example, they maintained that tobacco smoking was not established as a cause of lung cancer until both association studies and mechanistic studies had been evaluated (2007: 162-3). (Phyllis Illari later provided some clarification of the commitments of this Evidential Pluralism (2011: Mechanistic evidence: Disambiguating the Russo-Williamson thesis, *International Studies in the Philosophy of Science*, 25(2): 139–57). Brendan Clarke et al. give a more recent defence of this Evidential Pluralism in medicine (2014: Mechanisms and the evidence hierarchy, *Topoi*, 33: 339–60).)

An initial worry is that this Evidential Pluralism represents something of a step backwards in the evolution of evidence appraisal in medicine. Indeed, mistakes have been made in medicine by relying upon mechanistic studies, because the mechanisms at play were often more complicated than was acknowledged. This sort of worry has been pressed by Miriam Solomon (2015: *Making Medical Knowledge*, Oxford). However, these mistakes were arguably the result of relying upon mechanistic studies *alone*. The Evidential Pluralist proposal is to rely upon both association studies and mechanistic studies. The idea is that the limitations of mechanistic studies are compensated for by the strengths of association studies, and the limitations of association studies are compensated for by the strengths of mechanistic studies. (For more on this issue, see Daniel Auker-Howlett and Michael Wilde (2019: Reinforced reasoning in medicine, *Journal of Evaluation in Clinical Practice*, 26: 458–64).)

Another worry is that Evidential Pluralism is just not feasible in contemporary medicine; it is often hard enough to evaluate association studies, let alone evaluate association studies *alongside* mechanistic studies. However, Veli-Pekka Parkkinen et al. put forward guidelines for implementing Evidential Pluralism in medicine in a manageable way (Parkkinen et al. 2018: *Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures*, Springer). Moreover, Jon Williamson has recently argued that Evidential Pluralism is in fact feasible, since something at least very close to Evidential Pluralism is currently implemented by the International Agency for Research on Cancer (2020: The feasibility and malleability of EBM+, *Theoria*, 36(2): 191–209).

One objection to this Evidential Pluralism is that the distinction between association studies and mechanistic studies is too simplistic and clean-cut to do justice to the messy world of medicine. This sort of objection has been pressed by Raffaella Campaner and Maria Carla Galavotti (2012: Evidence and the assessment of causal relations in the health sciences, *International Studies in the Philosophy of Science*, 26(1): 27–45). Campaner and Galavotti argue both that the distinction between association and mechanistic studies misses out the important category of evidence from *manipulations*, and that evidence of association is often entangled with evidence of mechanisms.

Another objection is that the case studies cited in support of this Evidential Pluralism are controversial. In particular, Alex Broadbent has maintained that evaluating association studies alone was sufficient to establish that tobacco smoking was a cause of lung cancer (2011: Inferring causation in epidemiology: mechanisms, black boxes, and contrasts, in P. Illari, F. Russo, and J. Williamson (eds.), *Causality in the Sciences*, Oxford). Moreover, Jeremy Howick appeals to other case studies to argue similarly that sometimes evaluating associations studies alone is sufficient to establish causal claims in medicine (2011: Exposing the vanities—and a qualified defense—of mechanistic reasoning in health care decision making, *Philosophy of Science*, 78(5): 926–40).

Both of these objections have been at least indirectly addressed by Donald Gillies (2019: *Causality, Probability, and Medicine*, Routledge). He responds to the putative cases where associational studies alone established causal claims in medicine, and then argues that at least a modified version of Evidential Pluralism is consistent with the case study of tobacco smoking being established as a cause of lung cancer. Moreover, the notion of *interventional* or *manipulationist* evidence plays a prominent role in Gillies' discussion of Evidential Pluralism.

A residual worry may still remain about oversimplification. Bennett Holman argues that Evidential Pluralism is oversimplified in the sense that ignores the powerful economic forces at play in contemporary medicine (2019: Philosophers on drugs, *Synthese*, 196: 4363–90). And Mattia Andreoletti and David Teira argue that it is an oversimplification to provide only philosophical arguments in favour of Evidential Pluralism; there are costs associated with implementing any method for evaluating

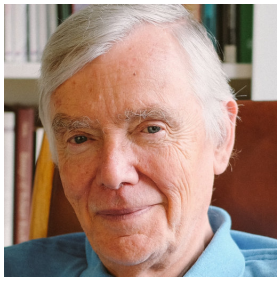
evidence in medicine, and Evidential Pluralism should be implemented only if it does best in the cost-benefit analysis comparing competing methods (2019: Rules versus standards: what are the costs of epistemic norms in drug regulation?, *Science, Technology, and Human Values*, 44(6): 1093–115).

So there remains some controversy surrounding Evidential Pluralism in the context of medicine. In light of this controversy, Jon Williamson has provided a sustained defence of Evidential Pluralism in medicine (2019: Establishing causal claims in medicine, *International Studies in the Philosophy of Science*, 32(1): 33–61).

Michael Wilde
Department of Philosophy and Centre for Reasoning
University of Kent

Evidential Pluralism and the Effectiveness of Treatments in Medicine

Evidential Pluralism applies where the evidence for a hypothesis h is of several different types. In medicine, there are two main types of evidence: (i) statistical evidence about human populations, and (ii) evidence of mechanism. Statistical evidence can be further divided into observational evidence such as is obtained in epidemiological surveys, and the evidence from clinical trials, which nowadays are nearly always randomized controlled trials or RCTs. The main thesis of Evidential Pluralism is that each type of evidence has both strengths and weaknesses, so that the best strategy for obtaining high empirical confirmation of h is to combine different types of evidence in such a way that the weaknesses of one type are cancelled out by the strengths of another type. In this way one can obtain higher confirmation of h than would be possible using just one type of evidence. This is the principle of *strength through combining* which was formulated by Phyllis Illari (2011: Mechanistic evidence: Disambiguating the Russo-Williamson thesis, *ISPS* 25(2): 139–57).



Let us now consider the case where our hypothesis h is that a particular treatment is an effective cure for some disease. Obviously we want to make sure that h is very well confirmed empirically before this particular treatment is introduced into medical practice. This case, however, appears *prima facie* to be a counter-example to Evidential Pluralism, because it looks as if the effectiveness of treatments can be tested out and confirmed or disconfirmed by just one type of evidence, namely that obtained by RCTs. If a treatment has been shown to be effective in a well-designed and conducted RCT, is there really any need to consider other types of evidence? It looks as if the evidence of RCTs alone is sufficient as a basis for deciding whether a treatment is effective.

Although this line of argument seems plausible, I now want to argue that it is wrong by considering a particular counter-example. This will show that, although RCTs definitely provide very strong evidence, they nevertheless do have some weaknesses, and these weaknesses can be overcome by considering evidence of another type, namely evidence of mechanism.

The example to be considered is a famous one. It concerns the three trials of streptomycin and other anti-tuberculosis agents which were carried out by the British Medical Research Council (MRC) in the period 1947-51. For reasons of space, the account given here of these trials is rather brief, but more details with references to the original papers is to be found in my book: Donald Gillies (2019: *Causality, Probability, and Medicine*, Routledge, 153–58). These trials are of considerable importance in the history of medicine, because they were among the first RCTs used in medicine, and they were one of the strong influences which led, quite rightly, to the increasing use of RCTs to test the efficacy of proposed medicines. However, if we examine these trials closely, we shall find that what they actually support is not the use of RCTs on their own, but rather *in conjunction with* evidence of mechanism.

Streptomycin was discovered in America in 1944 by Schatz, Bugie, and Waksman. It was shown that it strongly inhibited tubercle bacilli in vitro, and that it was also successful *in vivo* in treating experimental tubercular infections in guinea-pigs. The new antibiotic even produced some quite spectacular cures of patients suffering from tuberculosis. Streptomycin seemed a very promising candidate to be the long-sought cure for tuberculosis, and Austin Bradford Hill, who was a firm believer in the desirability of randomized controlled trials, managed to persuade the MRC to let him carry out a RCT to test Streptomycin's effectiveness. The first patients for it were recruited in January 1947.

The procedure was fairly straightforward. The patients were all between 15 and 30, and suffering from acute progressive bilateral pulmonary tuberculosis. Between January 1947 and September 1947, 109 patients had been accepted. 2 of these died in the preliminary observation week, and the remaining 107 were assigned randomly to either the control group C or the streptomycin group S. There were 52 in C, and 55 in S. The control group C received the standard treatment of the time, which was 6 months of bed-rest. The S group received, in addition, doses of streptomycin. The streptomycin was continued for four months, but the patients were observed for a further 2 months. So the trial was brought to a close for each patient after 6 months.

The results of the RCT showed that the S group did very considerably better than the C group, and so strongly supported the effectiveness of streptomycin. 51% of the S group showed considerable improvement as against only 8% of the C group. 7% of the S group died as against 27% of the C group. These differences are highly significant statistically.

In the light of such good results from the RCT, one might have expected that the MRC would have declared that treatment with streptomycin had been shown to work, and was to be recommended. Instead of giving such an endorsement of streptomycin therapy, however, the MRC concluded on a very cautious note. This caution proved to be amply justified. The same patients were investigated after 5 years, and it was then found that 58% of the S group had died as against 67% of the C group. The difference here is not statistically significant. What seems to have happened in the S group is that, after the encouraging initial improvement, many relapsed.

This example is an instance of a general problem with RCTs, which could be described as time limitations. Such trials have to come to an end after some time period t. Suppose the RCT shows that the treatment has produced a marked improvement by t, can we then be sure that this will not be followed by a relapse later on?

How can this problem be overcome? Well, those who conducted the streptomycin trial did seem to overcome the problem. They did foresee that the long-term results might not be so good as was suggested by the short-term improvements; and, for this reason, sounded a note of caution. How did they manage this? The answer is that they took account of evidence about the mechanism of the treatment.

Already by 1947 many researchers in the area had become aware that there might be a problem with streptomycin therapy. While some antibiotics such as penicillin could dispose of the pathogenic bacteria, which they targeted, in a week or two, streptomycin took many weeks, even months, to deal with a patient's tubercle bacilli. Now Darwinian evolution as applied to bacteriology strongly suggested that, in such a time period, strains of the tubercle bacillus might develop which would be resistant to streptomycin. Such resistant strains posed a very considerable threat to streptomycin therapy. They might well increase in numbers producing a relapse, and, in this new condition, a fresh treatment with streptomycin would obviously be useless.

Because of an awareness of this difficulty, those who carried out the streptomycin RCT, at the same time carried out an investigation into the mechanism of the treatment. They took samples of tubercle bacilli from the patients and tested them for resistance to streptomycin. At the beginning of the streptomycin treatment no bacilli from the patients were found to be resistant. However, by the end of the second month, 63% of the cases in the S group, which were examined, had developed resistance to streptomycin.

How was this problem to be overcome? The researchers had the ingenious idea of combining streptomycin with another anti-tubercle agent, which would dispose of the streptomycin resistant bacilli. The agent chosen was para-amino-salicylic acid or PAS. Subsequent RCTs showed that a combination of streptomycin with PAS had just as good results as streptomycin alone, while the accompanying investigation of the patient's bacilli showed that hardly any developed streptomycin resistance. In this way the first successful treatment for tuberculosis was discovered and the value of combination treatments was established. If the researchers had considered only the evidence of the original RCT, this would not have happened, which shows the great value of using different types of evidence.

Donald Gillies
Department of Science and Technology Studies
University College London

Evidential Pluralism and Sport and Exercise Science

It only takes a Google Scholar search to see how many practical disciplines, since the advent of Evidence Based Medicine (EBM), aim to employ an 'evidence-based' approach. One field where this is the case is Sport and Exercise Science (SES). SES investigates using biomechanics, physiology, nutrition, and psychology, to explain and intervene on things like fitness, health, and sports performance. The aim of adopting an evidence-based methodology in SES is to engage in Evidence Based Practice (EBP). By adopting methodologies from EBM in EBP, the hope is that practice will be informed by the best possible evidence, providing strong justification for practices.

Key to this is the adoption of evidence hierarchies. These supposedly rank the quality of evidence provided for a claim by study type. Borrowing from EBM, Knudson (2014: Proposing application of results in sport and exercise research reports, *Sports Biomechanics* 13(3), 195–203) suggests an evidence hierarchy typifying those employed in EBP (Figure 4). In this hierarchy, as in many others, we can see that Randomized Controlled Trials (RCTs) are ranked as providing strong evidence, where descriptive research, such as may be derived from mechanistic studies, is ranked as providing the lowest quality of evidence. As is emphasised by Ivarsson and Anderson (2016: What counts as "evidence" in evidence-based practice? Searching for some fire behind all the smoke, *JSPA* 7(1), 11–22), EBP 'privileges' evidence from RCTs. This often means that evidence from other sources, such as mechanistic studies, is dismissed. For this to be good practice, one needs to work on the assumption that RCTs can, and regularly do, provide strong evidence for causal claims. The hierarchies, the privileging of RCTs, and the fact that RCTs are treated as a 'gold standard' of evidence in medicine (Sackett et al. 1996: Evidence based medicine: what it is and what it isn't, *BMJ* 312, 71–72), show us that this assumption *is* often made.



Level	Research characteristics	Potential application	Qualifications/limitations
I	Reviews of prospective and implementation research (RCT)	Strong evidence	Individual response, barriers, risk/benefit
II	Prospective, implementation research	Preliminary evidence	Population
III	Experimental and retrospective research	Limited evidence	Prospective confirmation needed
IV	Descriptive research or technical note	Hypothesized evidence	Initial evidence needed

Note: Limitations associated with each level include the limitations of each higher level of evidence.

Figure 4: An evidence hierarchy specific to SES, reproduced from Knudson (2014).

I argue that this assumption is unjustified, however. Due to the nature of SES, RCTs will often be unable to meet the requirements needed to produce strong evidence. To provide strong evidence, RCTs need:

- large sample sizes in order to rule out chance correlation, and properly estimate and detect effect sizes,
- adequate placebo controlling in order to have a baseline outcome measure to compare other outcome measures to, and
- sufficient blinding in order to stop participant hunches impacting upon measured outcomes.

Sample sizes in SES will often be small, however. In the case of elite and niche sports, few potential participants exist. In other cases, such as exercise trials, it is hard to find participants willing to adhere to the long and involving interventions seen in sport. Placebo controlling is difficult, too. This is because the complexity of many sports interventions means it is near to impossible to find a placebo indistinguishable from the true

intervention, whilst also not including elements of the intervention being tested that could impact relevant outcome measures. It would be very difficult, for instance, to placebo control an exercise intervention, as is argued by Maddocks et al. (2016: Problematic placebos in physical therapy trials, *JECP* 22(4), 598–602). How would one make a participant think they were exercising, without impacting any of the outcome measures that exercise may influence? Potential solutions to the difficulties of adequately placebo controlling are also difficult to employ in SES:

- Dose response trials need large sample sizes and effective placebo controls to provide good evidence.
- Active control trials assume the active control being tested against already has established efficacy, which is difficult to show without already having a suitable placebo against which to test it first.

Finally, it is very difficult to sufficiently blind sport and exercise trials. Even if a participant does not know whether they are receiving an intervention or a placebo, it will often be easy for someone conducting or administering an intervention to guess. For instance, a masseuse will know they are giving a fake massage, and a coach will know if they are administering a sham training plan. This can lead to changes in application of interventions or placebos, and interpretation of results, which can affect outcome measures.

Where SES RCTs do not well fulfil these requirements, outcomes observed cannot be readily attributed to interventions or exposures under investigation. For instance, chance and hunches about what trial group one is in may be the true explanation of observed outcomes, not what is being tested. Accordingly, RCTs not fulfilling these requirements provide low quality evidence. In turn, whatever one thinks is needed in order to establish a causal claim, RCTs in SES will often be unable to provide strong evidence for it because observed outcomes cannot be readily attributed to interventions under investigation.

Evidential Pluralism both helps us to explain why RCTs failing to meet these requirements also fail to establish causal claims, and also helps to provide us with practical solutions to the problem of justifying practice.

Through the lens of Evidential Pluralism: we may observe a correlation in an RCT, but until we can establish that a mechanism exists to explain that the intervention caused it, we cannot attribute it to the intervention under investigation. As RCTs do not provide evidence for a mechanism by providing details of mechanisms, to provide good evidence for a mechanism they must be sufficiently rigorous to rule out other explanations. As RCTs in SES often fail to meet the requirements needed to rule out alternate explanations for observed outcomes, even where they indicate a correlation may exist, their failure to rule in a mechanism means they fail to provide strong evidence of causality.

As this is the case, EBP seems to be wrong in privileging evidence from RCTs. Practice could be better informed where we had evidence that also provided strong justification that mechanisms exist. As such, EBP ought to assess RCTs and mechanistic studies together when assessing causal claims and justifying practice.

William Levack-Payne
Department of Philosophy and Centre for Reasoning
University of Kent

Evidential Pluralism in the Social Sciences

Evidential Pluralism is a normative thesis concerning the epistemology of causation. It was first proposed by Russo and Williamson (2007: Interpreting causality in the health sciences, *ISPS* 21(2), 157–170), and further developed recently by Shan and Williamson (2021: Applying Evidential Pluralism to the social sciences, *EJPS* 11(4), 1-27). In a nutshell, Evidential Pluralism consists of two normative claims:

> (1) in order to establish a causal claim, one normally needs to have both evidence of correlation and evidence of mechanisms;
>
> (2) when assessing a causal claim, one ought to consider both association studies and mechanistic studies, where available.

Evidential Pluralism was originally introduced in the context of the health sciences and has been fruitfully applied to the biomedical sciences. However, the application of Evidential Pluralism to the social sciences has been controversial. For example, Weber (2009: How probabilistic causation can account for the use of mechanistic evidence, *ISPS* 23(3), 277–295) contends that Evidential Pluralism is 'correct' in the context of the social sciences, while Reiss (2009: Causation in the Social Sciences: Evidence, Inference, and Purpose, *PoSS* 39(1), 20–40) is sceptical of the application of Evidential Pluralism to the social sciences.

A major concern arises from the standard way of conceiving of the methodology of the social sciences, which focusses on a division between the quantitative tradition and qualitative tradition. For example, in sociology, quantitative researchers focus on statistical models and analyses and usually neglect the need to develop sociological models that mirror social mechanisms. In contrast, social theorists are mainly concerned with their concepts and theoretical frameworks and pay insufficient attention to the significance of quantitative findings. In political science, there has also been a methodological divide between the quantitative research approach and qualitative research approach. Such a methodological parallel pervades causal inquiry. When talking about causal analysis, social scientists tend to focus on looking for one type of evidence. For example, it is not unusual for political scientists to make within-case causal inferences by employing process-tracing methods to identify a mechanism. In other words, it seems to many that political scientists do not need any evidence of correlation when they establish single-case causal claims (e.g., the causes of the Russian Revolution). In a similar vein, Claveau (2012: The Russo-Williamson theses in the social sciences: Causal inference drawing on two types of evidence, *SHPSC* 43(4), 806–813) argues that economists can establish causal claims without evidence of correlation.



However, this is not quite right. As Shan and Williamson (2021: Applying Evidential Pluralism to the social sciences, *EJPS* 11(4), 1-27) have argued, those seeming counterexamples are not genuine counterexamples: process-tracing studies in political science typically assume some established correlations, while in Claveau's case, the relevant causal claims are

not established due to a lack of evidence of mechanisms and of correlation. Good social science research does accord well with the basic idea of Evidential Pluralism: social scientists tend to take both association studies and mechanistic studies into consideration when they assess causal claims.

A clear case is the study of socioeconomic status and health status. Social scientists have noticed that there is a strong association between socioeconomic status and health status. For example, lower socioeconomic status is associated with the 14 major causes of death in the International Classification of Diseases. In addition, lower socioeconomic status is shown to be associated with lower life expectancy, higher overall mortality rates, and higher rates of infant and perinatal mortality. However, it is debatable whether socioeconomic status is a cause of health status. Sceptics typically argue that socioeconomic status is a placeholder variable for real causes of diseases that have not yet been identified.

Even for some sociologists who argue for the causal relationship between socioeconomic status and health, a strong and pervasive association between socioeconomic status and health merely provides a description of the social pattern of disease. It is widely accepted that in order to establish the causal claim that socioeconomic status is a cause of disease, one has to establish the existence of some mechanisms as well as a correlation. As Link and Phelan (1995: Social Conditions As Fundamental Causes of Disease, *JHSB* Extra, 80–94) suggest, it is necessary to identify 'the direction of causation between social conditions and health and the mechanisms that explain observed associations' for the purpose of 'establishing a causal role for social factors'.

With their collaborators, Phelan and Link (2010: Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications, *JHSB* Sup, 28–40) have identified a variety of mechanisms linking socioeconomic status to health status. It is shown that persons of higher socioeconomic status possess a wide range of resources, including money, knowledge, power and beneficial social connections, which shape health-enhancing behaviours (such as getting flu jabs, eating fruits and vegetables, and exercising regularly) and access to broad contexts that are associated with risk and protective factors of health. For example, those who have lower status jobs more commonly have job strain (i.e., a combination of high job demands and low decision latitude), which is associated with coronary heart disease; people with lower socioeconomic status are more likely to smoke and be overweight, which lead to various health problems; and people with lower socioeconomic status experience greater residential crowding and noise, which is linked to poorer long-term memory and to reading deficits.

The debate over socioeconomic status and health thus illustrates that sociologists take both correlation and mechanism into account when they try to establish or assess a causal claim. The proponents of the theory of fundamental causes maintain that socioeconomic status is a fundamental cause of health status on the grounds that both the correlation and the mechanisms are established, while opponents challenge the causal claim by questioning the mechanism hypotheses. That both sides of the debate focus on evidence of correlation and evidence of mechanisms shows that Evidential Pluralism captures the structure of causal analysis in the social sciences.

Yafeng Shan
Department of Philosophy and Centre for Reasoning
University of Kent

Evidential Pluralism in Social Policy

In 2007, Russo and Williamson put forward a form of evidential pluralism that argues, among other things, that a causal claim can be established only if it can be established that there is a difference-making relationship between the cause and the effect, and that there is a mechanism linking the cause and the effect that is responsible for that difference-making relationship (Russo and Williamson 2007: Interpreting causality in the health sciences, *ISPS* 21(2), 157–170). The applicability of Evidential Pluralism to biomedical research and health policies has provoked a lot of debate (see for instance Clarke et al. 2014: Mechanisms and the Evidence Hierarchy, *Topoi* 33, 339–360; Williamson 2019: Establishing Causal Claims in Medicine, *ISPS* 32(1), 33–61; Parkkinen et al. 2018: Evaluating Evidence of Mechanisms in Medicine, *Springer* 33, 339–360). In particular, Parkkinen et al. developed methods for evaluating mechanistic studies alongside association studies in medicine. In the social policy domain, however, the debate is yet to be rigorously shaped.



If we look at the UK What Works Centres (WWCs) and similar evidence-based policy centres that support government to develop policy, programmes and services, it is evident that the main approach to explore 'what works' is the use of difference-making studies, in particular randomised controlled trials. A closer look at these centres, however, can lead to an interesting observation: evidence of difference-making relationships is frequently combined with evidence of mechanisms, but different terminologies and a lack of methodological discussion make it difficult to recognise its use.

Let's consider, for instance, a typical evaluation of an evidence-based intervention. Such an evaluation is very likely to include a robust RCT, often called an 'impact evaluation', that helps to collect evidence of the difference-making relationship between the intervention and the outcome(s) of interest. An assumption often made by UK What Works Centres and clearing houses is that robust RCTs can support causal claims by ruling out the risk of confounding, therefore leading to the conclusion that the difference-making relationship between the cause and the effect is due to the presence of a mechanism linking them. In other words, if we consider Figure 1, RCTs would support causal claims via routes C_1 and C_2, and evidence directly supporting a difference-making relationship would also indirectly support the presence of a mechanism.

In social policy, as in the social sciences, the use of RCTs has been challenged (Morris et al. 2016: The importance of specifying and studying causal mechanisms in school-based randomised controlled trials: lessons from two studies of cross-age peer tutoring, *Educational Research and Evaluation* 22, 339–360), and the debate has often been framed around a call for evaluation of logic models, or process evaluations.

In social policy, a logic model is defined as a graphical model that shows how an intervention leads to specific outputs, which in turn contribute to short- and long-term outcomes. Logic models are generally tested through process evaluations (also called implementation and process evaluations, IPEs) that through a mixed-methods approach collect and analyse evidence of the mechanisms and processes that are thought to trigger change in outcomes.

One of the WWCs that have been more explicit about the use of evidence of mechanisms in assessing interventions is the Education Endowment Foundation (EEF). EEF recently considered that identifying 'what works' through RCTs alone is not enough to ensure complex interventions will improve outcomes in the population. Hence, EEF argued, it is also important to ask how interventions work, and determine why they do or do not work, for whom and under what conditions they work. Implementation and process evaluation (IPE)—EEF claimed—'can help us to answer these important questions by providing researchers with theoretical, methodological and analytical tools that enable insights into the processes and mechanisms underpinning the impact (or lack thereof) of educational interventions' (EEF 2016: Implementation and process evaluation (IPE) for interventions ineducation settings: An introductory handbook, pp. 35–6).

There is at least one case where Evidential Pluralism (i.e., the combination of difference-making evidence from RCTs and mechanistic evidence from process evaluations) is explicitly discussed in social policy: the need for extrapolation. There is a general agreement that it is only by considering mechanistic evidence, and confirming the logic model, that policymakers can understand whether the outcomes obtained in a given context can be obtained, with the same intervention, in a new context.

When it comes to the typical 'what works' question, however, it is unclear what role, if any, process evaluations and the evidence gathered through these studies should have. Recently, some approaches have emerged that partly take into account evidence of mechanisms to establish the effectiveness of interventions. A good example is the 'EMMIE' framework developed by the What Works Centre for Crime Reduction for systematic reviews of evidence, which is focused on 5 factors: Effect size, Mechanism, Moderator, Implementation and Economics (Thornton et al., 2019: On the development and application of EMMIE: insights from the What Works Centre for Crime Reduction, *An International Journal of Research and Policy* 29(3), 266–282). As discussed by Shan and Williamson (2021: Applying Evidential Pluralism to the social sciences, *EJPS* 11(96)), this framework is based not on Evidential Pluralism, but on the realist evaluation approach, which among other things, rejects the experimental methodology that underpins RCTs and certain other kinds of difference-making study.

More work needs to be done to ensure that social policy approaches are based on coherent philosophical foundations. The call for logic models and process evaluations, and considerations of mechanistic evidence in reviews that explore interventions' effectiveness, can be seen as an important step in the right direction.

Virginia Ghiara
Department of Philosophy and Centre for Reasoning
University of Kent

Evidential Pluralism and Econometric Modelling

Evidential Pluralism is an epistemological and methodological thesis according to which, in order to establish a causal claim, one need both evidence of correlation and evidence of mechanism.

Evidential Pluralism was originally formulated in the context of the health sciences (Russo & Williamson 2007: Interpreting causality in the health sciences, *ISPS* 21(2), 157–170), and large part of the philosophical discussion since then remained within the medical domain, broadly construed. Only very recently have we seen strong interest in Evidential Pluralism in the context of the social sciences (Shan & Williamson 2021: Applying Evidential Pluralism to the social sciences, *EJPS*, in press). However, this is a renewed interest, as some attention to the social science domain had been given a few years back. Claveau (2012: The Russo-Williamson theses in the social sciences: Causal inference drawing on two types of evidence, *SHPSC* 43(4), 806–813), for instance, challenged the core idea of Evidential Pluralism by arguing that, in a particular case study he examined, scientists did not make any use of evidence of correlation. Shan & Williamson offer a defence of Evidential Pluralism against Claveau's argument.

I want to highlight here another piece of work on Evidential Pluralism in the social sciences, and specifically in econometric modelling. Together with Alessio Moneta, I looked at modelling in econometrics, and we examined the conditions under which we can establish causal claims (Russo & Moneta 2014: Causal models and evidential pluralism in econometrics, *JEM* 21, 54–76). Our strategy was to unfold the practice of econometric modelling, walking step by step through model building and model testing and distinguishing between associational and causal models, which I had already introduced in previous work, to shed light on causal modelling in the social sciences more generally (Russo 2009: *Causality and causal modelling in the social sciences*, Springer; 2011: Correlational data, causal hypotheses, and validity, *JGPS* 42, 85–107). Moneta and I noted that causal models are 'augmented' statistical models, because they incorporate important causal information. In a sense, Moneta and I agreed with the motto 'no causes in, no causes out', and we tried to explain, in the context of econometric modelling, how causes get in, and how they get out. Specifically, when we get to 'augment' a statistical model (which is only associational), there is a lot of information about causes and mechanisms that enter this augmentation. Part of this information is encapsulated into causal assumptions proper, such as temporal priority of the causes. But another part of this information is given by background knowledge, which in turn contains information about institutional mechanisms, theoretical knowledge, etc. In joint work with Michel Mouchart and Guillaume Wunsch, I also advanced the view that (structural) models represent mechanisms, via the recursive decomposition (Mouchart & Russo 2011: Causal explanation: Recursive decompositions and mechanisms, in *Causality in the sciences*, OUP, 317–337; Mouchart, Russo & Wunsch 2010: Inferring causal relations by modelling structures, *Statistica* LXX, 411–432).



One challenge that the social sciences face—and econometrics is no exception in this respect—is that a same conclusion may be supported by very different mechanisms. This is because social groups and social phenomena do not obey rigid, deterministic laws, and outcomes may be realized in multiple ways. In our view, this was not an objection against Evidential Pluralism per se, or against the maturity of the social sciences and econometrics, but rather an argument for the inherent *model-dependence* of causality. In a similar vein, we did not aim to use Evidential Pluralism as a fool-proof method, or the magic bullet solution to establish causal relations in a world that is notoriously messy. We instead acknowledged the inherent *fallibility* of econometric modelling, and therefore the need to explicitly discuss how to move from associational to causal models, and what would eventually ground the final outcome of the modelling exercise.

In our mind, the (conceptual) distinction between associational and causal models would help understand controversies. We examined a famous debate on money demand in the UK, between Friedman and Schwartz (1982: *Monetary trends in the United States and United Kingdom*, UCP) on the one side and Hendry and Ericsson (1985: Assertion without empirical basis, *International Finance Discussion Papers* 270; 1991: An econometric analysis of UK money demand in monetary trends, *American Economic Review* 81, 8–38) on the other side. The debate is usually framed as being about the results of the different studies. Instead, in our view, one should look at the entire modelling strategies and see why the two camps come to different conclusions. In this case, our diagnosis of the controversy is that there were quite some differences in the modelling of the data, as well as background and theoretical assumptions, that explain the divergence of opinions. Some of these can be linked to questions about evidence of correlation, and others to questions about evidence of mechanism.

There a number of questions that would be worthy of attention in the context of modelling and evidence, in the social sciences and elsewhere. For instance, often we are not precise enough about whether the model *generates* evidence of mechanism, or whether it *represents* it. It is plausible to think that some models fare better with the former task, and others fare better with the latter task.

Federica Russo
Department of Philosophy
University of Amsterdam

What counts as evidence for a mechanism? And why?

Evidential Pluralism—or better, 'evidential dualism'—insists that good confirmation that *C* causes/caused/will cause *E* generally requires evidence that *C* makes/made/will make a difference to *E* and evidence that there's a mechanism connecting *C* and *E*. My topic here is the latter: what counts as evidence for a connecting mechanism? This is a question we seldom see explicit answers to.



Here I take mechanisms to be step-by-step processes connecting *C* and *E* in which each step contributes to producing the next, and I suppose that any cause cited at a step is, as JL Mackie (1965: Causes and Conditions, *APQ* 2, 245–264) argued, an INUS condition for the next—an Insufficient but Necessary part of an Unnecessary but Sufficient condition for a contribution.

Here are some of the evidence-types we use for supporting causal claims across the sciences and in everyday life, including some of the Bradford-Hill criteria. How compelling such evidence is all told depends case-by-case on how much there is, of what kinds and how sure we can be of the evidence claims.

1. Evidence that helps eliminate alternatives.
2. Evidence about the character of the effect. Does *E* occur at the time, in the manner and of the size to be expected had *C* contributed to it?
3. Evidence about the size of other factors affecting *E*.
4. Symptoms of causation, like side effects to be expected had *C* operated to produce *E*.
5. Presence of support factors (moderator/interactive variables) that need to be in place for *C* to contribute to *E*.
6. Presence of expectable intermediate steps (mediator variables).

Here's a caricature illustration. Imagine that yesterday I consumed a harmful poison. Luckily I realised I'd done so and thereafter swallowed a strong emetic. I vomited violently and have subsequently not suffered serious symptoms of poisoning. I praise the emetic: it saved me! What evidence can support that?

- *Elimination of alternatives.* There are low survival rates with this poison. So it's not likely my survival was spontaneous. And there's nothing special about me that would otherwise explain my survival. I don't have an exceptional body mass, I hadn't been getting slowly acclimatised by earlier smaller doses, I didn't take an antidote, etc.
- *Presence of required support factors.* The emetic was swallowed before too much poison was absorbed from the stomach.
- *Presence of necessary intermediate steps.* I vomited.
- *Presence of symptoms of the causes acting to produce the effect.* There was much poison in the vomit, which is a clear side effect of the emetic's being responsible for my survival.
- *Characteristics of the effect.* The amount of poison in the vomit was measured and compared with the amount I had consumed. I suffered just the effects of remaining amount of poison, and the timing of the effect and size were just right.

I claim that these are all evidence about a connecting mechanism. What *justifies* that they are? They are evidence about a connecting mechanism because they provide information about the 'situation-specific causal equation model' (SCEM) that represents the mechanism and its pre- and post-history. A SCEM is a set of potential outcome equations (POEs) modelling how *C* brings about *E*, step-by-step in the targeted situation. POEs describe a full set of causes for the effect. They serve as a

ground for counting RCT results as evidence of causation (see, e.g., Deaton 2020: Randomization in the Tropics Revisited: A Theme and Eleven Variations, NBER Working Paper w27600). Since RCTs are touted as the best kind of difference-making evidence, POEs are crucial for both difference-making and mechanistic evidence.

To build a SCEM, start with E. What should C have led to at the previous stage to produce E? Call that E_{-1}. Then what support factors are necessary for E_{-1} to produce E? Represent their net effect by α_{-1}. Next, what other factors will be in place at the penultimate stage that affect E? Represent their net effect by W_{-1}. This gives a POE like this:

$$E = \alpha_{-1}E_{-1} + W_{-1}.$$

Note that each variable should be time indexed. Complete the model by working backwards constructing a POE for each step till we reach C.

But there's more. Consider the support factors represented by the α_i. These are themselves both causes and effects. Knowing about the causes of causes of an effect is a clue to whether the causes will occur and thus to whether the effect can be expected. Their causal history can be expressed in POEs added to the SCEM. The factors that do not interact with E_i (represented by W_i) but also affect E_{i+1} also have causal histories, represented in POEs that can be added. We may also want to include equations in which $E, E_{-1}, E_{-2}, \ldots$ figure as a cause since seeing that their effects obtain gives evidence that they occurred. We can include as much or as little of the causal histories of various variables in the SCEM as we find useful.

Why construct a SCEM? Because this is what Nature does. What causes what is not arbitrary—at least not where we can expect to make reasonable predictions, explanations and evaluations. There's a system to how Nature operates. Some factors *can* affect E in this situation and some *cannot*. All those that can affect E appear in Nature's own POEs for E.

Supposing that, look back to my list of evidence types. It is readily apparent how they map onto the SCEM. For instance, support factors tell us about the α_i. If any of these in any of the equations is missing, the chain from C to E is broken. The W_i affect the size of the E_i. Estimating if E_i is the right size to produce E_{i+1} given α_i tells you about W_i, and the reverse. And so forth. When C causes E, Nature has her own model of the connecting mechanism and of its pre- and post-history. That's why facts about features in a SCEM provide us with evidence for a mechanism between C and E.

Nancy Cartwright
Department of Philosophy
Durham University and UCSD

Evidential Pluralism in Cognitive Science

In the context of their discussion of the health and biomedical sciences, Russo and Williamson (2007: Interpreting causality in the health sciences, *ISPS* 21(2), 157–170) put forward the following claim about the epistemology of causation:

> In order to establish that A is a cause of B in medicine one normally needs to establish two things. First, that A and B are suitably correlated—typically, that A and B are probabilistically dependent, conditional on B's other known causes. Second, that there is some underlying mechanism linking A and B that can account for the difference that A makes to B.

This claim has come to be known as the Russo-Williamson Thesis and, more generally, as *Evidential Pluralism*.

Evidential pluralism has already found good support in the health and social sciences. But open questions remain about its applicability elsewhere. Here, I will suggest that Evidential Pluralism can be—and, in fact, has been—fruitfully applied in the cognitive sciences. To make this case, I will consider an example from cognitive neuroscience: Dehaene's (2009: *Reading in the brain*, New York) theory of reading.



Dehaene argues that "the brain contains fixed circuitry exquisitely attuned to reading" and that the functional activity of this cortical area is causally responsible for our capacity to recognise words and letters. The area in question is located in the left ventral occipito-temporal junction and is now commonly labelled with a functional designation that Dehaene himself coined: the visual word form area (VWFA).



Figure 1: fMRI of the visual word form area (VWFA) from Muayqil, T., Davies-Thompson, J., & Barton, J. J. (2015: Representation of visual symbols in the visual word processing network, *Neuropsychologia* 69, 232-241).

The idea is that the function of VWFA is causally responsible for certain behaviours (namely, letter/word recognition and reading), because patterns of activity in VWFA are correlated with some environmental parameters of relevance (e.g. the presence of letters/words in the visual field) and these patterns of activation play a causal role in the cognitive process that enables the organism to read by acting as a signal that informs the activities of downstream neural mechanisms.

(Note that cognitive neuroscientists typically speak of a brain area as having a particular function if that area contributes to the operation of the system of which it is a part by representing an element in the task-environment of the organism. There is no space here, however, to engage in a discussion about what it means to say that "x is a representation," where x is some pattern of neural activity.)

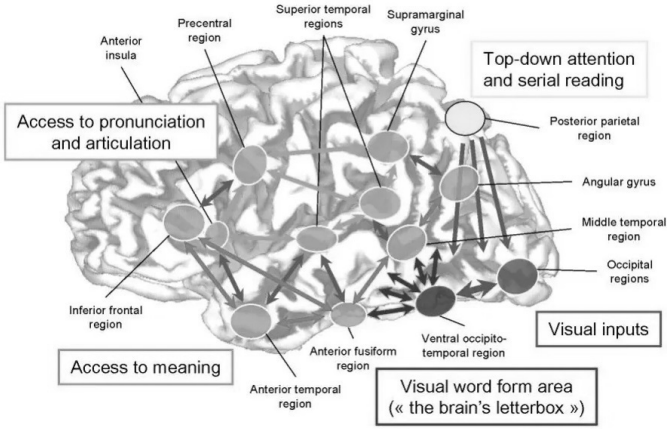


Figure 2: A modern vision of the cortical networks for reading from Dehaene (2009).

Thus, we have a causal claim: that patterns of activity in the VWFA (in response to certain environmental parameters) cause cognitive processes that enable organisms to recognise letters / words and, ultimately, to read. The open question, however, is how this causal claim is established. And it is here where Evidential Pluralism seems to be the causal epistemology at work. The reason is that Dehaene only feels able to put forward the aforementioned causal claim after establishing evidence of *both* mechanisms and correlation.

In this case, the evidence of correlation is plentiful. For example, there is evidence that in normal literate subjects, VWFA is differentially responsive to written, but not spoken words (Dehaene and Cohen 2007: Cultural recycling of cortical maps *Neuron* 56(2), 384-398); that in blind subjects, the region is differentially responsive to words presented in Braille, but not to tactile control stimuli (Reich et al. 2011: A ventral visual stream reading center independent of visual experience, *Current Biology* 21(5), 363-368); and that lesions to VWFA appear to result in pure alexia, a condition in which formerly literate subjects cannot understand written words, despite being able to understand and produce verbal speech at roughly normal levels of competency (Gaillard et al. 2006: Direct intracranial, FMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading, *Neuron* 50(2), 191-204).

In each of these examples, we find that the functioning of VWFA (or not) makes a difference to our capacity to recognise letters/words and to read. This just is the kind of evidence of correlation that Russo and Williamson had in mind.

But Dehaene also appeals to evidence of mechanisms; specifically, a mechanism for reading that links the functional activities of VWFA to the functional activities of other areas of the brain (see Figure 2) (Dehaene et al. 2002: The visual word form area: a prelexical representation of visual words in the fusiform gyrus, *Neuroreport* 13(3), 321-325). The idea here is that:

> The left occipitotemporal "letterbox" [(e.g. VWFA)] identifies the visual form of letter strings. It then distributes this invariant visual information to numerous regions, spread over the left hemisphere, that encode word meaning, sound pattern, and articulation. [...] Learning to read thus consists in developing an efficient interconnection between visual areas and language areas. All connections are bidirectional.

Of course, the mechanistic details here remain vague, because the detailed organisation of these brain areas is not yet fully known and cortical connectivity is highly complex. But there is good evidence of the organisation and activities of different regions of the brain (e.g., posterior parietal region, occipital regions) linking VWFA to our capacity to recognise letters/words and to read goes beyond mere association (Cohen et al. 2002: Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area, *Brain* 125(5), 1054-1069).

We find, therefore, that Evidential Pluralism at least accords with the causal inferences of working cognitive neuroscientists like Dehaene. Still, further research is required to substantiate the normative claim that Evidential Pluralism applies to *every* case of causal discovery in cognitive science. Moreover, further research is required to assess what prescriptions Evidential Pluralism makes as to the way in which causal inference in the cognitive sciences—and cognitive neuroscience, in particular—is carried out.

Samuel D. Taylor
Department of Philosophy and Centre for Reasoning
University of Kent

Evidential Pluralism and Explainable AI

What is the relationship between machine learning and the philosophy of science? I have previously argued that it is one of *dynamic interaction* (2004: A dynamic interaction between machine learning and the philosophy of science, *Minds and Machines* 14(4), 539–549; 2010: The philosophy of science and its relation to machine learning, in *Scientific Data Mining and Knowledge Discovery*, Springer, pp. 77–89). A dynamic interaction is a mutually beneficial interaction between two autonomous fields (Gillies & Zheng 2001: Dynamic interactions with the philosophy of mathematics, *Theoria* 16(3), 437–459).



A dynamic interaction can provide a fertile source of progress of the two interacting fields. This progress can sometimes be attributable to particular loci of interaction. Here I would like to suggest that a connection between Evidential Pluralism in the philosophy of science and explainable AI in machine learning may provide a fruitful locus of interaction.

Explainable AI aims to produce machine learning systems that are easier to assess. There is a story—perhaps apocryphal—of a neural network that was trained to recognise and classify tanks but which ended up basing its classifications on incidental features of the images presented to it: on the weather conditions instead of features of the tanks themselves. Another system for screening applications of prospective medical students was found to discriminate against women and applicants with non-European names. The hope is that if an AI system were to provide explanations of its decisions, it would be easier to pick up on such deficiencies. More recently, Uber's autonomous vehicle that killed a lady crossing a road in Tempe, Arizona, in 2018, kept misclassifying her and failed to determine her trajectory. In this case, it probably wouldn't have been helpful to provide explanations in real time, but the killing

might have been avoided if the system and its training process were more interpretable.

Thus a distinction is sometimes drawn between *black box AI*, where often even its designers can't explain why the AI system arrived at a decision, *explainable AI*, whose results can be explained to human users, and *interpretable AI*, whose methods can be directly interpreted and understood by human users. Complex neural nets and evolutionary algorithms are often taken to be black box systems, while at the other end of the spectrum, simple decision trees and rule-based systems are usually taken to be interpretable. The suggestion is that explainable and interpretable AI provide more confidence in reliability, robustness and fairness, and may satisfy regulatory demands for the right to explanation for certain decisions. Note that in the case of explainable AI, explanations are often provided by a second AI system. This leads to a kind of regress problem: how can we be certain that this second system is reliable, robust and fair? Interpretable AI may avoid this kind of regress.



A connection can be forged between Evidential Pluralism and explainable AI. On the one hand, Evidential Pluralism is a very natural approach from the perspective of explainable AI. With reference to Fig. 1, C-channel confirmation is 'black-box' and non-robust, while M-channel confirmation elucidates mechanisms and mechanisms are explanatory. Establishing causation by means of both channels offers what might be called 'explainable' causal discovery. This can yield confidence that the observed correlation isn't spurious.

On the other hand, explainable AI is a very natural approach from the perspective of Evidential Pluralism. The use of an AI system can be viewed as an intervention to achieve a certain goal. We can ask whether the intervention is effective: is it a cause of the goal being achieved? Evidential Pluralism directly applies to this causal question. Association studies are standardly used to test an AI system. These are prone to biases, such as over-fitting the study population from which training and test data are drawn. Explanations often tell us about the relevant mechanisms, which can help to rule out these biases. Association studies remain crucial, however, because the mechanism of the AI system is usually too complex to be able to predict a correlation from it. Appealing to both allows us to assess whether the system works in the study population.

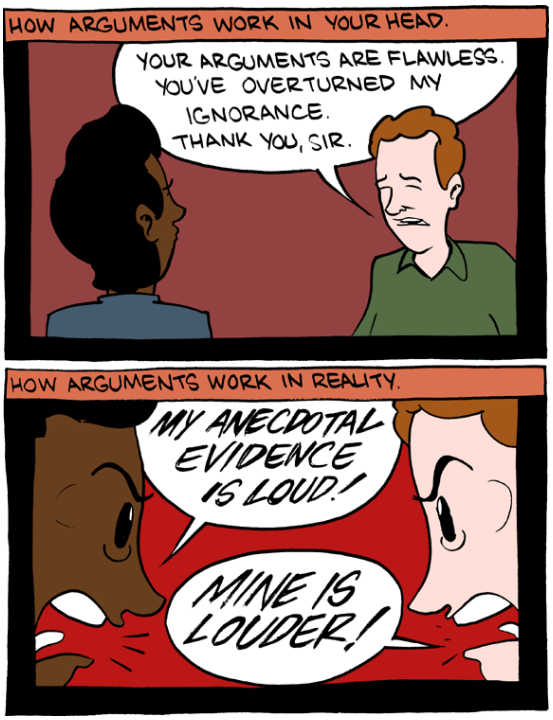
The connection between Evidential Pluralism and explainable AI has the potential to provide a fruitful point of interaction because each can suggest new directions for the other.

Explainable AI suggests the following development of Evidential Pluralism. For a particular causal claim, one might be able establish the existence of an appropriate mechanism that can account for an observed correlation, and even establish the details of the mechanism, without being able to understand how the mechanism explains the correlation. For example, as Samuel Taylor points out above, one may have strong evidence of the cognitive mechanisms by which the visual word form area of the brain causes certain reading abilities, without being able to understand exactly how the mechanisms give rise to these abilities. While such an understanding may not be required to establish causality, explainable AI suggests that understanding is important in order to provide an intervention's stakeholders with confidence that the intervention will work and is fair. Thus, in cases where intervention assessment includes an element of stakeholder interaction, there may be a need for 'interpretable Evidential Pluralism' that goes beyond Evidential Pluralism as currently formulated.

On the other hand, Evidential Pluralism may be of benefit to explainable AI in the following ways. For Evidential Pluralism, mechanism details are helpful for determining both internal and external validity, for intervention refinement, and, as Gillies argues above, for the proper design and interpretation of association studies. All these four tasks are crucial to machine learning. We need to know whether an AI system works in the target environment as well as the test environment, we need to be able to refine and improve the system, and we need to design and interpret good tests of the system's reliability. Evidential Pluralism thus offers an integrated approach that has the potential to meet several of the needs of explainable AI.

Moreover, lessons can be drawn for explainable AI from the use of Evidential Pluralism in evidence based medicine. The initial vision for EBM was for all clinicians to use EBM to evaluate interventions. This soon turned out to be unrealistic, because the evaluation process is so time-consuming and complex. These days, trusted expert committees are charged with evaluating interventions and less is expected of clinicians. Perhaps the current vision for explainable and interpretable AI is also unrealistic—i.e., the aspiration that key stakeholders will be in a position to evaluate AI systems. If so, we need trusted expert committees for AI evaluation. This would amount to a radical change in the oversight of AI systems.

JON WILLIAMSON
Department of Philosophy and Centre for Reasoning
University of Kent



Calls for Papers

Reasoning with Inconsistent, Incomplete, and Uncertain Knowledge: special issue of *IEEE Intelligent Systems*, deadline 30 December.
The Epistemology of Computer Simulations: special issue of *Balkan Journal of Philosophy*, deadline 30 December.

Events

December

TCiPoS: Thick Concepts in the Philosophy of Science, Hannover, Germany, 3–4 December.
IPS: Interdisciplinarity and Philosophy of Science, online, 6–8 December.
HTA: Current & Future Statistical Issues & Challenges in Health Technology Assessment, online, 7 December.

Courses and Programmes

Courses

CiE: Computability in Europe 2021: Connecting with Computability Tutorials, 5–9 July.

Programmes

MA in Reasoning, Analysis and Modelling: University of Milan, Italy.
APhil: MA/PhD in Analytic Philosophy, University of Barcelona.
Master Programme: MA in Pure and Applied Logic, University of Barcelona.
Doctoral Programme in Philosophy: Language, Mind and Practice, Department of Philosophy, University of Zurich, Switzerland.
Doctoral Programme in Philosophy: Department of Philosophy, University of Milan, Italy.
LogiCS: Joint doctoral program on Logical Methods in Computer Science, TU Wien, TU Graz, and JKU Linz, Austria.
HPSM: MA in the History and Philosophy of Science and Medicine, Durham University.
Master Programme: in Statistics, University College Dublin.
LoPhiSC: Master in Logic, Philosophy of Science and Epistemology, Pantheon-Sorbonne University (Paris 1) and Paris-Sorbonne University (Paris 4).
Master Programme: in Artificial Intelligence, Radboud University Nijmegen, the Netherlands.
Master Programme: Philosophy and Economics, Institute of Philosophy, University of Bayreuth.
MA in Cognitive Science: School of Politics, International Studies and Philosophy, Queen's University Belfast.
MA in Logic and the Philosophy of Mathematics: Department of Philosophy, University of Bristol.
MA Programmes: in Philosophy of Science, University of Leeds.
MA in Logic and Philosophy of Science: Faculty of Philosophy, Philosophy of Science and Study of Religion, LMU Munich.
MA in Logic and Theory of Science: Department of Logic of the Eotvos Lorand University, Budapest, Hungary.
MA in Metaphysics, Language, and Mind: Department of Philosophy, University of Liverpool.
MA in Mind, Brain and Learning: Westminster Institute of Education, Oxford Brookes University.
MA in Philosophy: by research, Tilburg University.
MA in Philosophy, Science and Society: TiLPS, Tilburg University.
MA in Philosophy of Biological and Cognitive Sciences: Department of Philosophy, University of Bristol.
MA in Rhetoric: School of Journalism, Media and Communication, University of Central Lancashire.
MA programmes: in Philosophy of Language and Linguistics, and Philosophy of Mind and Psychology, University of Birmingham.
MRes in Methods and Practices of Philosophical Research: Northern Institute of Philosophy, University of Aberdeen.
MSc in Applied Statistics: Department of Economics, Mathematics and Statistics, Birkbeck, University of London.
MSc in Applied Statistics and Datamining: School of Mathematics and Statistics, University of St Andrews.
MSc in Artificial Intelligence: Faculty of Engineering, University of Leeds.
MSc in Cognitive & Decision Sciences: Psychology, University College London.
MSc in Cognitive Systems: Language, Learning, and Reasoning, University of Potsdam.
MSc in Cognitive Science: University of Osnabrück, Germany.
MSc in Cognitive Psychology/Neuropsychology: School of Psychology, University of Kent.
MSc in Logic: Institute for Logic, Language and Computation, University of Amsterdam.
MSc in Mind, Language & Embodied Cognition: School of Philosophy, Psychology and Language Sciences, University of Edinburgh.
MSc in Philosophy of Science, Technology and Society: University of Twente, The Netherlands.
MRes in Cognitive Science and Humanities: Language, Communication and Organization: Institute for Logic, Cognition, Language, and Information, University of the Basque Country (Donostia San Sebastián).
Open Mind: International School of Advanced Studies in Cognitive Sciences, University of Bucharest.
Research Master in Philosophy and Economics: Erasmus University Rotterdam, The Netherlands.

Jobs and Studentships

Studentships

Doctoral Programme in Philosophy: Language, Mind and Practice, Department of Philosophy, University of Zurich, Switzerland.
LogiCS: Joint doctoral program on Logical Methods in Computer Science, TU Wien, TU Graz, and JKU Linz, Austria.

Jobs

Lectureship: in Theoretical Philosophy, University of Leiden, deadline 10 December.
PhD position: in Knowledge Representation and Reasoning, University of Luxembourg, deadline 10 December.

PhD position: in Uncertainty Quantification for Precision Medicine, University of Exeter, deadline 10 January.
Post-doc: in History of Philosophy of Science, Tilburg University, deadline 14 January.

