**Causal Discovery and the Problem of Psychological Interventions**

PSA 2018, Seattle

Markus Eronen

University of Groningen

m.i.eronen@rug.nl

**Abstract**

Finding causes is a central goal in psychological research. In this paper, I argue that the search for psychological causes faces great obstacles, drawing from the interventionist theory of causation. First, psychological interventions are likely to be both fat-handed and soft, and there are currently no conceptual tools for making causal inferences based on such interventions. Second, holding possible confounders fixed seems to be realistically possible only at the group level, but group-level findings do not allow inferences to individual-level causal relationships. I also consider the implications of these problems, as well as possible ways forward for psychological research.

**1. Introduction**

A key objective in psychological research is to distinguish causal relationships from mere correlations (Kendler and Campbell 2009; Pearl 2009; Shadish and Sullivan 2012). For example, psychologists want to know whether having negative thoughts is a cause of anxiety instead of just being correlated with it: If the relationship is causal, then the two are not just spuriously hanging together, and intervening on negative thinking is actually one way of reducing anxiety in patients suffering from anxiety disorders. However, to what extent is it actually possible to find psychological causes? In this paper, I will seek an answer this question from the perspective of state-of-the-art philosophy of science.

In philosophy of science, the standard approach to causal discovery is currently interventionism, which is a very general and powerful framework that provides an account of the features of causal relationships, what distinguishes them from mere correlations, and what kind of knowledge is needed to infer them (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009, Woodward 2003, 2015b; Woodward & Hitchcock 2003). Interventionism has its roots in Directed Acyclic Graphs (DAGs), also known as causal Bayes nets, which are graphical representations of causal relationships based on conditional independence relations (Spirtes, Glymour and Scheines 2000; Pearl 2000, 2009). More recently, James Woodward has developed interventionism into a full-blown philosophical account of causation, which has become popular in philosophy and the sciences. Several authors have also argued that interventionism adequately captures the role of causal thinking and reasoning in psychological research (Campbell 2007; Kendler and Campbell 2009; Rescorla forthcoming; Woodward 2008).

Based on interventionism, I will argue in this paper that the discovery of psychological causes faces great obstacles. This is due to problems in performing psychological interventions and deriving interventionist causal knowledge from psychological data.[[1]](#footnote-1) Importantly, my focus is not on the existence or possibility of psychological causation, but on the *discovery* of psychological causes, which is a topic that has so far received little attention in philosophy.[[2]](#footnote-2) Although I rely on interventionism, my arguments are based on rather general principles of causal inference and reasoning in science, and will thus apply to any other theory of causation that does justice to such principles.

The focus in this paper will be on the discovery *individual-level* (or within-subject)causes, not *population-level* (or between-subjects) causes. The first refers to causal relationships that hold for a particular individual: for example, John’s negative thoughts cause John’s problems of concentration. The latter refers to causal relationships that obtain in the population as a whole: for example, negative thoughts cause problems of concentration in a population of university students. It widely thought that ultimate goal of causal inference is to find individual-level causes, and that a population-level causal relationship should be seen as just an average of individual-level causal relationships (Holland 1986): For example, the causal relationship between negative thoughts and problems of concentration in a population of university students is only interesting insofar as it *also* applies to at least some of the individual students in the population.[[3]](#footnote-3) Thus, in this paper I will discuss population-level causal relationships only when they are relevant to discovering individual-level causes.

Importantly, the distinction between population-level and individual-level causation is different from the distinction between type and token causation, even though the two distinctions are sometimes mixed up in the philosophical literature (see also Illari & Russo 2014, ch. 5). Token causation refers to causation between two actual events, whereas type causation refers to causal relationships that hold more generally. Individual-level causes can be either type causes or token causes. An example of an individual and type causal relationship would be that John’s pessimistic thoughts cause John’s problems of concentration: This is a general relationship between two variables, and not a relationship between two actual events. An example of an individual and token causal relationship would be that John’s pessimistic thoughts before the exam on Friday at 2 pm caused his problems of concentration in the exam. As interventionism is a type-level theory of causation, and the aim of psychological research is primarily to discover regularities, not explanations to particular events, in this paper I will only discuss the discovery of type (individual) causes.

The structure of this paper is as follows. I will start by giving a brief introduction to interventionism, and then turn to problems of interventionist causal inference in psychology: First, to problems related to psychological interventions (section 2), and then to problems arising from the requirement to “hold fixed” possible confounders (section 3). After this, I will consider the possibility of the inferring psychological causes without interventions (section 4). In the last section, I discuss ways forward and various implications that my arguments have for psychology and its philosophy.

**2.****Interventionism**

Interventionism is a theory of causation that aims at elucidating the role of causal thinking in science, and defining a notion of causation that captures the difference between causal relationships and mere correlations (Woodward 2003). Thus, the goal of interventionism is to provide a methodologically fruitful account of causation, and *not* to reduce causation to non-causal notions or analyse the metaphysical nature of causation (Woodward 2015b). In a nutshell, interventionist causation is defined as follows:

X is a cause of Y (in variable set **V**) if and only if it is possible to *intervene* on X to change Y when all other variables (in **V**) that are not on the path from X to Y are *held fixed* to some value (Woodward 2003).

Thus, in order to establish that X is a cause of Y, we need evidence that there is some way of intervening on X that results in a change in Y, when off-path variables are held fixed.[[4]](#footnote-4) Importantly, it is not necessary to actually perform an intervention: What is necessary is knowledge on what *would* happen if we *were* to make the right kind of intervention.

The notion of an intervention plays a fundamental role in the account, and is very specifically defined. Here is a concise description of the four conditions that an intervention has to satisfy (Woodward 2003).

Variable *I* is an intervention variable for X with respect to Y if and only if:

(I1) *I* *causes* the change in X;

(I2) The change in X is *entirely* due to *I* and not any other factors;

(I3) *I* is not a cause of Y, or any cause of Y that is not on the path from X to Y;

(I4) *I* is *uncorrelated* with any causesof Y that are not on the path from X to Y.

The rationale behind these conditions is that if the intervention does not satisfy them, then one is not warranted to conclude that the change in Y was (only) due to the intervention on X. Thus, in simpler terms, the intervention should be such that it changes the value of the target variable *X* in such a way that the change in *Y* is *only* due to the change in *X* and not any other influences (Woodward 2015b). For example, if the intervention is correlated with some other cause of Y, say Z, that is not on the path from X to Y (violating I4), then the change in Y may have been (partly) due to Z, and not just due to X. Following standard terminology in the literature, I will call interventions that satisfy the criteria I1-I4 *ideal* interventions. I will now go through various problems in performing ideal interventions in psychology, starting from problems related to conditions I2 and I3 (section 3), and then turn to problems related to I4 and the “holding fixed” part of the definition of causation (section 4).

**3. Psychological interventions**

Before discussing psychological interventions, an important distinction needs to be made: The distinction between relationships where (1) the cause is *non-psychological*, and the *effect* is psychological, and (2)where the *cause* (and possibly also the effect) is *psychological*.[[5]](#footnote-5) A large proportion (perhaps the majority) of experiments in psychology involve relationships of the first kind: The intervention targets a non-psychological variable (X) such as medication vs. placebo, therapy regime vs. no therapy, or distressing vs. neutral video, and the psychological effect of the manipulation of this non-psychological variable is tracked. In other words, the putative causal relation is between a non-psychological cause variable (X) and a psychological effect variable (Y). In these cases, it is possible to do (nearly) ideal interventions on the putative cause variable (X) by ensuring that the change in X was caused (only) by the intervention, that the intervention did not change Y directly, and that it was uncorrelated with other causes of Y. It is of course far from trivial to make sure that these conditions were satisfied, but as the variables intervened upon are non-psychological, making the right kinds of interventions is in principle not more difficult than in other fields. As regards the psychological effect variable (Y), there is no need to intervene on it; it is enough to measure the change in Y (which, again, is far from trivial, but faces just the usual problems in psychological measurement, which will be discussed below). The fact that many psychological experiments involve this kind of causal relationships may have contributed to the recent optimism on the prospects of interventionist causal inference in psychology.

However, psychological research also often concerns relationships of the second kind, that is, relationships where the *cause* is psychological. This is, for example, the case when the aim is to uncover psychological mechanisms that explain cognition and behavior (e.g., Bechtel 2008, Piccinini & Craver 2011), or to find networks of causally interacting emotions or symptoms (e.g., Borsboom & Cramer 2013). The reason why these relationships are crucially different from relationships of the first kind is that now the variable intervened upon is psychological, so the conditions on interventions now have to be applied to psychological variables.

Ideal interventions on psychological variables are rarely if ever possible. One reason for this has been extensively discussed by John Campbell (2007): Psychological interventions seem to be “soft”, meaning that the value of the target variable *X* is not completely determined by the intervention (Eberhardt & Scheines 2007; see also Kendler and Campbell 2009; Korb and Nyberg 2006). In other words, the intervention does not “cut off” all causal arrows ending at *X*. As a non-psychological example, when studying shopping behaviour during one month by intervening on income, an ideal intervention would fully determine the exact income that subjects have that month, whereas simply giving the subjects an *extra* 5000€ would count as a soft intervention (Eberhardt & Scheines 2007). Similarly, if we intervene on John’s psychological variable *alertness* by shouting “WATCH OUT!”, this does not completely cut off the causal contribution of other psychological variables that may influence John’s *alertness*, but merely adds something on top of those causal contributions (Campbell 2007). As most or all interventions on psychological variables are likely to be soft, Campbell proposes that we should simply allow such soft interventions in the context of psychology. Campbell argues that these kind of interventions can still be informative and indicative of causal relationships (Campbell 2007), and this conclusion is supported by independent work on soft interventions in the causal modelling literature (e.g., Eberhardt & Scheines 2007; Korb and Nyberg 2006).

However, the problem of psychological interventions is not solved by allowing for soft interventions. There is a further, equally important reason why interventions on psychological variables are problematic: Psychological interventions typically *change several variables simultaneously*. For example, suppose we wanted to find out whether *pessimistic thoughts* cause *problems in concentration*. In order to do this, we would have to find out what would happen to *problems in concentration* if we were to intervene just on *pessimistic thoughts* without perturbing other psychological states with the intervention. However, how could we intervene on *pessimistic thoughts* without changing, for example, *depressive mood* or *feelings of guilt*? As an actual scientific example, consider a network of psychological variables that includes, among others, the items *alert*, *happy*, and *excited* (Pe et al. 2015). How could we intervene on just one of those variables without changing the others?

One reason why performing “surgical” interventions that only change one psychological variable is so difficult is that there is no straightforward way of manipulating or changing the values of psychological variables (as in, for example, electrical circuits). Interventions in psychology have to be done, for example, through verbal information (as in the example of John above) or through visual/auditory stimuli, and such manipulations are not precise enough to manipulate just one psychological variable. Also state-of-the-art neuroscientific methods such as Transcranial Magnetic Stimulation affect relatively large areas of the brain, and are not suited for intervening on specific psychological variables. Currently, and in the foreseeable future, there is no realistic way of intervening on a psychological variable without at the same time perturbing some other psychological variables.

Thus, it is likely that most or even all psychological interventions do not just change the target variable X, but also some other variable(s) in the system. In the causal modelling literature, interventions of this kind have been dubbed *fat-handed*[[6]](#footnote-6) interventions (Baumgartner and Gebharter 2016; Eberhardt & Scheines 2007; Scheines 2005). For example, an intervention on pessimistic thoughts that also immediately changes depressive mood is fat-handed. Fat-handed interventions have been recently discussed in philosophy of science, but mainly in the context of mental causation and supervenience (e.g., Baumgartner and Gebharter 2016, Romero 2015), and the fact that psychological interventions are likely to be systematically fat-handed (for reasons unrelated to supervenience) has not yet received attention.

An additional complication is that it is difficult check what a psychological intervention precisely changed, and to what extent it was fat-handed (and soft). In fields such as biology or physics there are usually several independent ways of measuring a variable: for example, temperature can be measured with mercury thermometers or radiation thermometers, and the firing rate of a neuron can be measured with microelectrodes or patch clamps. However, measurements of psychological variables, such as emotions or thoughts, are based on self-reports, and there is no further independent way of verifying that these reports are correct. Moreover, only a limited number of psychological variables can be measured at a given time point, so an intervention may always have unforeseen effects on unmeasured variables.

Why are fat-handed interventions so problematic for interventionist causal inference? The reason becomes clear when looking at condition I3: The intervention should not change any variable Z that is on a causal pathway that leads to Y (except, of course, those variables that are on the path between X and Y). If the causal structure of the system under study is known, as well as the changes that the intervention causes, then this condition can sometimes be satisfied even the intervention was fat-handed. However, in the context of intervening on psychological variables, neither the causal structure nor the exact effects of the interventions are known. Thus, when the intervention is fat-handed, it is not known whether I3 is satisfied or not, and in many cases it is likely to be violated. In other words, we cannot assume that the intervention was an unconfounded manipulation of X with respect to Y, and cannot conclude that X is a cause of Y.

**4. The Problem of “Holding Fixed”**

The next problem that I will discuss is related to the last part of the definition of interventionist causation: X is a cause of Y (in variable set **V**) if and only if it is possible to intervene on X to change Y *when all other variables (in* ***V****) that are not on the path from X to Y are held fixed to some valu*e. The motivation for this requirement is to make sure that the change in Y is really due to the change X, and not due to some other cause of Y. To a large extent, this is just another way of stating what is already expressed in the definition of an intervention, in conditions I3 and I4: The intervention should not be confounded by any cause of Y that is not on the path between X and Y.[[7]](#footnote-7) In the previous section, we saw that fat-handed interventions pose a challenge for satisfying this condition. However, as I will now show, it is problematic in psychology also for more general reasons.

In psychology, it is impossible to hold psychological variables fixed in any concrete way: We cannot “freeze” mental states, or ask an individual to hold her thoughts constant. Thus, the same effect has to be achieved indirectly, and the gold standard for this is Randomized Controlled Trials (RCTs) (Woodward 2003, 2008). RCTs have their origin in medicine, but are widely used in psychology and the social sciences as well (Clarke et al. 2014; Shadish, Cook and Campbell 2002; Shadish and Sullivan 2012). The basic idea of RCTs is to conduct a trial with two groups, the test group and the control group, which are as similar to one another as possible, but the test group receives the experimental manipulation and the control group does not. If the groups are large enough and the randomization is done correctly, any differences between the groups should be only due to the experimental manipulation. If everything goes well, this in effect amounts to “holding fixed” all off-path variables.

However, this methodology has an important limitation that has been overlooked in the literature on interventionism. As the effect of “holding fixed” is based on the difference between the groups as wholes, it only applies at the level of the group, and not at the level of individuals. For this reason, results of RCTs hold for the study population as a whole, but not necessarily for particular individuals in the population (cf. Borsboom 2005, Molenaar & Campbell 2009). For example, if we discover that pessimistic thoughts are causally related to problems of concentration in the population under study, it does not follow that this causal relationship holds in John, Mary, or any other specific individual in the population. This is related to the “fundamental problem of causal inference” (Holland 1986): Each individual in the experiment can belong to only one of the two groups (control or test group), and therefore cannot act as a “control” for herself, so only an average causal effect can be estimated. What this implies for causal inference in psychology is that when a causal relationship is discovered through an RCT, we cannot infer that this relationships holds for any specific individual in the population (see also Illari & Russo 2014, ch. 5).

This does not mean that the population-level findings based on RCTs are uninformative or useless. The point is rather that we currently have no understanding of when, to what extent and under what circumstances they also apply to the individuals in the population. This of course applies also to other fields where RCTs are used, such the biomedical sciences. Indeed, especially in the context of personalized medicine, the fact that RCTs are as such not enough to establish individual-level causal relationships has recently become a matter of discussion (e.g., de Leon 2012).

It might be tempting to simply look at the data more closely and find those individuals for whom the intervention on X actually corresponded with a change in Y. However, it would be a mistake to conclude that in those individuals the change in Y was caused by X. It might very well have been caused by some other cause of Y, as possible confounders were only held fixed at the group level, not at the individual level.[[8]](#footnote-8) Thus, in RCTs possible confounders can only be held fixed at the group level, and this does not warrant causal inferences that apply to specific individuals. This is further limitation to interventionist causal inference in psychology.

**5. Finding psychological causes without interventions**

One possible response to the concerns raised in the previous two sections is that interventionism does not require that interventions are actually performed: As briefly mentioned in section 2, what is necessary is to know what *would* happen if we *were* to perform the right kinds of interventions. In other words, in order to establish that X is a cause of Y, it is enough to know that if we *were* to intervene on X with respect to Y (while holding off-path variables fixed), then Y *would* change. For example, it is beyond doubt that the gravitation of the moon causes the tides, even though no one has ever intervened on the gravitation of the moon to see what happens to the tides, and such an intervention would be practically impossible (Woodward 2003). Similarly, it could be argued that even though it is practically impossible to do (ideal) interventions on psychological variables, the knowledge on the effects of interventions could be derived in some other way. Let us thus consider to what extent this could be possible.

The state-of-the-art method for deriving (interventionist) causal knowledge when data on interventions is not available is *Directed Acyclic Graphs (DAGs)*, which were briefly mentioned in the introduction (see also Malinsky & Danks 2018, Pearl 2000, Spirtes, Glymour and Scheines 2000, Spirtes & Zhang 2016). Causal discovery algorithms based on DAGs take purely observational data as input, and based on conditional independence relations, find the causal graph that best fits the data. In principle, these algorithms can be used for psychological data, with the aim of discovering causal relationships between psychological variables.

However, even though these algorithms do not require experimental data, they do require data from which conditional independence relations can be reliably drawn, and they (implicitly) assume that the variables that are modelled are independently and surgically manipulable (Malinsky & Danks 2018). In contrast, as should be clear from the above discussion, measurements of psychological variables typically come with a great deal of uncertainly, and it is not clear to what extent they can be independently manipulated. Moreover, causal discovery algorithms standardly assume *causal sufficiency*, that is, that there are no unmeasured common causes that could affect the causal structure (Malinsky & Danks 2018; Spirtes & Zhang 2016). The reason for this is that if two or more variables in the variable set have unmeasured common causes, then the inferences concerning the causal relationships between those variables will be either incorrect or inconclusive. However, missing common causes is likely the norm rather than the exception when it comes to psychological variables. For example, if the variable set consists of, say, 16 emotion variables, how likely is it that *all* relevant emotion variables have been included? And even if all emotion variables that are common causes to other emotion variables are included, is it plausible to assume that there are no further cognitive or biological variables that could be common causes to some of the emotion variables? As similar questions can be asked for any context involving psychological variables, causal sufficiency is a very unrealistic assumption for psychological variable sets.

For these reasons, psychological data sets are rather ill-suited for causal discovery algorithms, and these algorithms cannot be treated as reliable guides to interventionist causal knowledge in psychology. It is likely that the problems of psychological interventions discussed in the previous sections are not just practical problems in carrying out interventions, but reflect the immense complexity of the system under study (the human mind-brain), and therefore cannot be circumvented by just using non-experimental data (see, however, section 7 for a different approach).

**6. Psychological interventions: A summary**

To summarize, what I have argued so far is that interventionist causal inference in psychology faces several obstacles: (1) Psychological interventions are typically *both* fat-handed *and* soft: They change several variables simultaneously, and do not completely determine the value(s) of the variable(s) intervened upon. It is not known to what extent such interventions give leverage for causal inference. (2) Due to the nature psychological measurement, the degree to which a psychological intervention was soft and fat-handed, or more generally, what the intervention in fact did, is difficult to reliably estimate. (3) Holding fixed possible confounders is only possible at the population level, not at the individual level, and it is not known under what conditions population-level causal relationships also apply to individuals. (4) Causal inference based on data without interventions requires assumptions that are unrealistic for psychological variable sets. Taken together, these issues amount to a formidable challenge for finding psychological causes.[[9]](#footnote-9)

**7. Discussion**

Although various metaphysical and conceptual issues related to psychological causation have been extensively discussed in philosophy of science, little attention has been paid to the *discovery* of psychological causes. In this paper, I have contributed to filling this lacuna, by discussing the search for psychological causes in the framework of the interventionist theory of causation. The upshot is that finding individual psychological causes faces daunting challenges. The problems in holding fixed confounders and performing interventions need to be taken into account when trying to establish a psychological causal relationship, or when making claims about psychological causes.

However, I do not want to argue that finding psychological causes is *impossible*, or that researchers should stop looking for psychological causes. Rather, my aim is to contribute to getting a better understanding of the limits of finding causes in psychology, and the challenges involved. This can also lead to positive insights regarding causal inference in psychology. One such insight is that more attention should be paid to *robust inference* or *triangulation*. Often when individual methods or sources of evidence are insufficient or unreliable, as is the case here, what is needed is a more holistic approach. A widespread (though not uncontroversial) idea in philosophy of science is that evidence from several independent sources can lead to a degree of confidence even if the sources are individually fallible and insufficient (Eronen 2015, Kuorikoski & Marchionni 2017, Munafo & Smith 2017, Wimsatt 1981, 1994/2007). For example, there is no single method or source of evidence that would be individually sufficient to establish that the anthropogenic increase in carbon dioxide is the cause for the rise in global temperature, but there is so much converging evidence from many independent sources that scientists are confident that this causal relationship exists. Similarly, evidence for a psychological causal relationship could be gathered from many independent sources: Several different (soft and fat-handed) interventions involving different variables, multilevel models based on time-series data, single-case observational studies, and so on.[[10]](#footnote-10) If they all point towards the same causal relationships, this may lead to a degree of confidence in the reality of that relationship. However, how this integration of evidence would exactly work, and whether it can actually lead to sufficient evidence for psychological causal relationships, are open questions.

A related point is that psychological research can also make substantive progress *without* establishing causal relationships. Often important discoveries in psychology have not been discoveries of causal relationships, but rather discoveries of robust *patterns* or *phenomena* (Haig 2012, Rozin 2001, Tabb and Schaffner 2017)*.* Consider, for example, the celebrated discovery that people often do not reason logically when making statistical predictions, but rely on shortcuts, for example, grossly overestimating the likelihood of dying in an earthquake or terror attack (Kahneman & Tversky 1973). In other words, when we reason statistically, we often rely on heuristics that lead to biases. The discovery of this phenomenon had nothing to with methods of causal inference (Kahneman and Tversky 1973), and its significance is not captured by describing causal relationships between variables. In fact, the causal mechanisms underlying the heuristics and biases of reasoning are still unknown. Similar examples abound in psychology: Consider, for example, groupthink or inattentional blindness. Of course, there are likely to be causal mechanisms that give rise to these phenomena, but the phenomena are highly relevant for theory and practice even when we know little or nothing about those underlying mechanisms (which is the current situation). This, in combination with the challenges discussed in this paper, suggests that (philosophy of) psychology might benefit from reconsidering the idea that discovering causal relationships is central for making progress in psychology.

Finally, one might wonder whether the problems I have discussed here are restricted to just psychology. Indeed, I believe that the arguments I have presented are more general, and apply to any other fields where there are similar problems with soft and fat-handed interventions and controlling for confounders. There is probably a continuum, where psychology is close to one end of the continuum, and at the other end we have fields where ideal interventions can be straightforwardly performed and variables can be easily held fixed, such as engineering science. Fields such as economics and political science are probably close to where psychology is, as they also face deep problems in making (ideal) interventions and measuring their effects. Same holds for neuroscience, at least cognitive neuroscience: The problems of soft and fat-handed interventions and holding variables fixed apply just as well to brain areas as to psychological variables (see also Northcott forthcoming). Thus, appreciating the challenges I have discussed here and considering possible reactions to them could also benefit many other fields besides psychology.

To conclude, I have argued in this paper that there are several serious obstacles to the discovery of psychological causes. As it is widely assumed in both psychology and its philosophy that the discovery of causes is a central goal, these obstacles need to be explicitly discussed, taken into account, and studied further.

**References**

Baumgartner, M. (2013). Rendering Interventionism and Non-Reductive Physicalism Compatible. *dialectica* 67: 1-27.

Baumgartner, M. (2018). The Inherent Empirical Underdetermination of Mental Causation. *Australasian Journal of Philosophy.*

Baumgartner, M and Gebharter, A. (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science* 67: 731-756.

Borsboom, Dennny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press

Borsboom, Denny and Anqelique O. Cramer. 2013. “Network analysis: an integrative approach to the structure of psychopathology.” *Annual review of clinical psychology* 9: 91-121.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203.

Campbell, John. 2007. “An interventionist approach to causation in psychology.” In: A. Gopnik & L. Schulz (eds.) *Causal Learning. Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 58–66.

Chirimuuta, Mazviita. Forthcoming. “Explanation in Computational Neuroscience: Causal and Non-causal.” *British Journal for the Philosophy of Science*. DOI:https://doi.org/10.1093/bjps/axw034

Clarke, B., Gillies, D., Illari, P, Russo, F., & Williamson, J. 2014. “Mechanisms and the evidence hierarchy.” *Topoi* 33: 339-360.

de Leon, J. (2012). Evidence-based medicine versus personalized medicine: are they enemies? *Journal of clinical psychopharmacology*, *32*(2), 153-164.

Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, *80*(5), 684-696.

Eberhardt, F. (2014). Direct causes and the trouble with soft interventions. *Erkenntnis*, *79*(4), 755-777.

Eberhardt, Frederick and Richard Scheines. 2007. “Interventions and causal inference.” *Philosophy of Science* 74: 981–995.

Eronen, Markus. Forthcoming. “Interventionism for the Intentional Stance: True Believers and Their Brains.” *Topoi*.

Hamaker, Ellen L. 2011. “Why researchers should think **“**within-person.**””** In M. R. Mehl, & T. A. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43**–**61). New York, NY: Guilford Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Kahneman, Daniel and Amos Tversky. 1973. “On the psychology of prediction.” *Psychological Review* 80: 237-251.

Kendler, Kenneth S. and John Campbell. 2009. Interventionist causal models in psychiatry: repositioning the mind-body problem. *Psychological Medicine* 39: 881-887.

Korb, K. B., & Nyberg, E. 2006. “The power of intervention.” *Minds and Machines* 16*:* 289-302.

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science, 83,* 227-247.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470.

Menzies, Peter. 2008. “The exclusion problem, the determination relation, and contrastive causation.” In J. Hohwy & J. Kallestrup (Eds.) *Being Reduced* (pp. 196-217). Oxford: Oxford University Press.

Molenaar, Peter and Cynthia Campbell. 2009. “The new person-specific paradigm in psychology.” *Current Directions in Psychological Science* 18: 112-117.

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature* 553, 399-401

Northcott, R. (forthcoming). Free will is not a testable hypothesis. *Erkenntnis*.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... & Kuppens, P. 2015. “Emotion-network density in major depressive disorder.” *Clinical Psychological Science*, *3*(2), 292-300.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, Judea. 2009. “Causal inference in statistics: An overview.” *Statistics surveys* 3: 96-146.

Pearl, Judea. 2014. “Comment: understanding simpson’s paradox.” *The American Statistician* 68: 8-13.

Peters, J. , Bühlmann, P. and Meinshausen, N. (2016), Causal inference by using invariant prediction: identification and confidence intervals. J. R. Stat. Soc. B, 78: 947-1012. doi:[10.1111/rssb.12167](https://doi.org/10.1111/rssb.12167)

Rescorla, Michael. Forthcoming. “An interventionist approach to psychological explanation.” *Synthese*.

Reutlinger, Alexander and Juha Saatsi (eds.). 2017. *Explanation Beyond Causation*. Oxford: Oxford University Press.

Romero, F. (2015). Why there isn’t inter-level causation in mechanisms. *Synthese*, *192*(11), 3731-3755.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2-14.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science*, *72*(5), 927-940.

Shadish W. R., Cook T. D. and Campbell D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton-Mifflin; Boston.

Shadish, W. R., & Sullivan, K. J. 2012. “Theories of causation in psychological science.” In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 23-52). Washington, DC: American Psychological Association.

Shapiro, Lawrence. 2010. “Lessons from causal exclusion.” *Philosophy and Phenomenological Research*, 81, 594-604.

Shapiro, Lawrence. 2012. “Mental manipulations and the problem of causal exclusion.” *Australasian Journal of Philosophy*, 90, 507–524.

Shapiro, Lawrence and Elliott Sober. 2007. “Epiphenomenalism: the dos and the don’ts.” In G. Wolters & P. Machamer (Eds.) *Thinking about causes: from Greek philosophy to modern physics* (pp. 235–264). Pittsburgh, PA: University of Pittsburgh Press.

Spirtes, Peter, Glymour, Clark and Richard Scheines. 2000. *Causation, prediction, and search*. New York: Springer.

Tabb, K., & Schaffner, K. F. (2017). Causal pathways, random walks and tortuous paths: Moving from the descriptive to the etiological in psychiatry. In: Kendler, K. S., & Parnas, J. (Eds.) *Philosophical Issues in Psychiatry IV: Nosology* (pp. 342-360). Oxford: Oxford University Press.

Weinberger, Naftali. 2015. “If intelligence is a cause, it is a within-subjects cause.” *Theory & Psychology*, 25(3), 346-361.

Woodward, James. 2003. *Making things happen. A theory of causal explanation.* Oxford: Oxford University Press.

Woodward, James. 2008. “Mental causation and neural mechanisms.” In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: new essays on reduction, explanation, and causation*. Oxford: Oxford University Press: 218-262

Woodward, James. 2015a. “Interventionism and causal exclusion.” *Philosophy and Phenomenological Research* 91, 303-347.

Woodward, James. 2015b. “Methodology, ontology, and interventionism.” *Synthese* 192, 3577-3599.

Woodward, James & Christopher Hitchcock. 2003. “Explanatory Generalizations, Part I: A Counterfactual Account.” *Noûs* 37(1): 1–24.

1. See Eberhardt (2013; 2014) for different (and domain-independent) problems for interventionist causal discovery. [↑](#footnote-ref-1)
2. There is an extensive debate on the question whether interventionism vindicates non-reductive psychological causation by providing a solution the causal exclusion problem (e.g., Baumgartner 2009, Eronen 2012, Raatikainen 2010, Woodward 2015). I will sidestep this debate here, as my focus is not on the existence of non-reductive psychological causation, but on the discovery of psychological causes, be they reducible or not. [↑](#footnote-ref-2)
3. It has been argued that population-level (between-persons) causal relationships can also be real without applying to any individual (Borsboom, Mellenbergh, and van Heerden 2003). However, also those who believe in these kind of population-level causes agree that discovering individual causes is an important goal as well. [↑](#footnote-ref-3)
4. More precisely, this is the definition for a *contributing* cause. X is a *direct* cause of Y if and only if it is possible to intervene on X to change Y when all other variables (in **V**) are heldfixed to some value (Woodward 2003). Thus, the definition of a contributing cause allows there to be other variables on the causal path between X and Y, whereas the definition of a direct cause does not. This does not reflect any substantive metaphysical distinction, as the question whether X is a direct or contributing cause is relative to what variables are included in the variable set. Importantly, notion of a contributing cause is *not* relative to a variable set in any strong sense – if X is a cause of Y in some variable set, then X will be a cause of Y in all variable sets where X and Y appear (Woodward 2008b). This is because the definition of an intervention is not relativized to a variable set. [↑](#footnote-ref-4)
5. The line between psychological and non-psychological variables is likely to be blurry. However, for the present purposes it is not crucial where exactly the line should be drawn: My arguments apply to cases where it is clear that the cause variable is psychological (such as the examples in the main text), and such cases abound in psychological research. [↑](#footnote-ref-5)
6. According to Scheines (2005), this term was coined by Kevin Kelly. [↑](#footnote-ref-6)
7. In recent publications, Woodward often gives a shorter definition of causation that does not include the “holding fixed” part, for example: “X causes Y if and only if under some interventions on X (and possibly other variables) the value of Y changes” (Woodward 2015). This is understandable, as the definition of intervention already contains conditions I3 and I4, which effectively imply holding fixed potential causes of Y that are correlated with the intervention and are not on the path from X to Y. However, there are also good reasons why the full definition has to include the second component as well. For example, consider a situation where we intervene on X with respect to Y, and Y changes, but this change is fully due to a change in variable Z, which is a cause of Y that is *uncorrelated* with the intervention variable. In this situation, without the “holding fixed” requirement we would falsely conclude that X is a cause of Y. [↑](#footnote-ref-7)
8. Would it be possible for a causal relationship to hold at the population level, but not for any individual in the population? Probably not, if the relationship is genuine: Weinberger (2015) has argued that there has to be at least *one* individual in the population for whom the relationships holds. However, in the context of discovery, it is possible that a causal *finding* at the population level is just an artefact of heterogeneous causal structures at the individual level, and therefore does not apply to any individual in the population. [↑](#footnote-ref-8)
9. Baumgartner (2009, 2012, 2018) has argued that mental-to-physical supervenience makes it impossible to satisfy the Woodwardian conditions on interventions, and that if interventionism is modified to accommodate supervenience relationships (as in Woodward 2015), the result is that any causal structure with a psychological cause becomes empirically indistinguishable from a corresponding structure where the psychological variable is epiphenomenal. If this reasoning is correct, it leads to a further (albeit more theoretical) problem for interventionist causal inference: Any empirical evidence for a causal relationships with a psychological cause is equally strong evidence for a corresponding epiphenomenal structure, and it is not clear which structure should be preferred and on what grounds. [↑](#footnote-ref-9)
10. See also Peters, Bühlmann, & Meinshausen (2016), who present a formal model for inferring causal relationships based on their stability under different kinds of (non-ideal) interventions. [↑](#footnote-ref-10)