In International Journal of Epidemiology, 2014

From Handles to Interventions: Commentary on R.G. Collingwood, "The So-Called Idea of Causation"

I.

R. G. Collingwood's paper, " On the So-Called Idea of Causation" ranges over a large number of topics connected with the notion of cause, but probably the portion of the paper that has attracted the most interest, both within and outside philosophy, is his discussion of what he calls "sense II" of "cause" (one of three senses of this notion that Collingwood distinguishes). He writes:

> In sense II, … the word cause expresses an idea relative to human action: the action in this case is an action intended to control, not other human beings, but things in "nature", or "physical" things. In this sense, the "cause" of an event in nature is the handle, so to speak, by which we can manipulate it….
>
> This sense of the word may be defined as follows. *A cause is an event or state of things which (i) it is in our power to produce or prevent, and (ii) by producing or preventing which we can produce or prevent that whose cause it is said to be*. (p.89)

(Italics in original. I have inserted the roman numerals (i) and (ii) in order to refer back to the separate clauses of this definition.)

He adds

> The search for causes in sense II is "natural science" in that sense of the phrase in which natural science is what Aristotle calls a "practical science," valued not for its truth pure and simple but for its utility, for the "power over nature" (p.90)

This is a statement of what is sometimes called a manipulability (or manipulationist) notion of causation, the basic idea being that a cause is a factor such that if it were possible to appropriately manipulate it, there would be an associated change in the effect. Broadly similar ideas have been defended by a number of other philosophers, including Gasking, 1955, von Wright, 1971, and Price, 1991, as well other researchers described below. Similar ideas are also common in many other disciplines, including econometrics, and statistics, as well as medicine.

The past several decades have seen the development of accounts of causation (and associated accounts of testing for causal relationships) that can be thought of as broadly manipulationist in spirit but that go well beyond the formulation provided by Collingwood—these include Spirtes, Glymour and Scheines, 2000, Pearl, 2000, and, in the philosophical literature, Woodward, 2003. A central concept in these accounts is the notion of an "intervention" (see below) and the associated idea that causal claims can be illuminated in various ways by construing them as claims about the outcomes of

hypothetical experiments[1]. In what follows I first discuss some aspects of Collingwood's paper and then try to place the paper in the context of these more recent contributions.

Manipulability accounts of causation have an obvious appeal. For one thing they seem to explain (at least in part) why we ought to care about whether a relationship is causal (rather than non-causal or "merely correlational"). Since we care about manipulation and control, particularly in "practical" subjects such as engineering and medicine, there is no mystery about why we also should care about the distinction between relationships that can be exploited for purposes of manipulation and control and those that cannot, and this is exactly the distinction between causal and non-causal relationships, according to the manipulationist conception. In addition, manipulationist conceptions of causation provide a transparent connection between the content of causal claims and one important way of testing such claims – via experimentation. Because manipulationist accounts tell us to interpret causal claims as claims about what would happen to some candidate effect if the candidate cause were to be experimentally manipulated, they make it obvious why, in order to test a causal claim, it is relevant to actually perform the indicated experiment if one can do that.

Despite this, philosophical assessment of manipulability accounts has been largely negative, for several reasons[2]. First, such accounts have struck many philosophers as leading to a conception of causation that is unacceptably anthropomorphic and "subjective". This is because, at least as often formulated, such accounts seem to make the truth or "meaning" of causal claims dependent on facts about what human beings can or can't do. This is true of Collingwood's formulation, for example, since it explicitly makes whether *X* causes *Y* (in Collingwood's sense II) dependent on whether it is in the power of human beings to manipulate *X* (cf. clause (i) from the quotation above); in cases in which a causal claim involves an unmanipulable cause ("the gravitational attraction of the moon causes the motion of the tides"), the logic of Collingwood's position requires him to hold that this involves a completely distinct notion of cause (his sense III.) For a variety of reasons this seems implausible—see my discussion below. A more satisfactory position is that causal relationships involving causes not currently manipulable by humans involve the same sense or concept of cause as relationships involving manipulable causes.

A second source of concern about manipulability accounts is that they have seemed to many philosophers to be objectionably circular, in the sense that they attempt to use notions that are causal in character (like "manipulation") to explicate what is for one factor to cause another. If (as seems plausible) "manipulate" means something like "cause to change", how can we use the notion of manipulation to get any independent purchase on the notion of causation? Many philosophers have supposed that a satisfactory account of causation must be "reductive" in the sense of explaining this notion in terms of concepts that do not themselves carry causal commitments, such as

---

[1] The potential outcomes framework developed by Rubin, 1974 also makes use of the idea that causal claims can be understood as claims about the outcomes of potential experiments, and thus is "manipulationist" in spirit but does not make explicit use of the notion of an intervention. Pearl, 2000 shows how the potential outcomes framework can be captured within a structural equations framework, given the notion of an intervention.

[2] See, e.g., the discussion in Hausman, 1998, 86ff.

"correlation". Interestingly, non-philosophers interested in causal inference, including those advocating manipulability theories do not share this concern about reduction, as will become apparent below.

A final concern is this: even if it is conceded (as suggested above) that manipulability accounts yield some insight into the special role of experimentation in establishing causal claims, what use are they in contexts in which we must rely on non-experimental evidence in establishing causal claims? What use is it to be told that if it were possible to manipulate *X* and *Y* would change under this manipulation, then *X* causes *Y*, if we cannot in fact manipulate *X*?

II.

I believe it is possible to provide satisfactory responses to all of these concerns, thus vindicating the basic thrust of the manipulability account. However, in order to do so, we must refine Collingwood's formulation. As a point of departure, consider the following difficulty, which is distinct from those mentioned above. Suppose that some factor *C* is a common cause of two correlated joint effects E_1 and E_2, and that there is no direct causal relationship between E_1 and E_2, themselves, as represented in the following directed graph.

$E_1 \leftarrow C \rightarrow E_2$

As an illustration, *C* might measure whether a subject smokes, E_1 might measure whether or not the subject has yellow fingers, and E_2 whether the subject develops lung cancer. Whether a subject has yellow fingers is correlated with whether the subject develops lung cancer but only because smoking acts as a common cause of both effects. Now suppose that one manipulates E_1 by manipulating *C*. Obviously E_2 will also change under this manipulation of E_1 so that if one applies the principle (suggested by the quotation from Collingwood above) that *X* causes *Y* whenever manipulation of *X* is associated with changes in *Y*, one will be led to the mistaken conclusion that E_1 causes E_2. What this shows is that not just any manipulation of a variable *X* should be regarded as an "appropriate" manipulation for the purposes of determining whether *X* causes *Y*. In recent literature, a more restricted, technical notion of manipulation has been introduced to avoid the difficulty just described. This is the notion of an "intervention" (cf. Spirtes, Glmour and Scheines, 2000, Pearl 2000, Woodward, 2003). Informally, an intervention on some candidate cause *X* with respect to a candidate effect *Y* is an "exogenous" change in *X* which is such that *Y* is affected, if at all, only through its relationship to *X* and not via some other causal route that does not go through *X*. Put slightly differently, an intervention on a candidate cause *X* with respect to a candidate effect *Y* should manipulate *X* in a way that is causally and statistically independent of all other causes of *Y* except those that lie directly on the causal route, if any, from *X* to *Y*. (Intuitively, these other "off-route" causes of *Y* are potential confounding factors that need to be controlled for.) In the example above, an intervention on E_1 would involve manipulating E_1 in a way that is independent of *C* and other factors that might potentially confound the relationship between E_1 and E_2 This might be accomplished by, for example, a randomized experiment in which subjects are assigned to conditions involving either yellow or non-yellow fingers (e.g. by dying non-yellow fingers yellow or by removing stains from

yellow fingers) in a way that is independent of whether the subject smokes. Of course we expect that under such a manipulation, there will be no association between finger color and the development of lung cancer and this shows that finger color does not cause lung cancer. (For more details, as well as a discussion of the representation of interventions by means of directed graphs and structural equation modelling, see Spirtes, Glymour and Scheines, 2000, Pearl, 2000, Woodward, 2003).

Given the notion of an intervention, we might replace Collingwood's version of a manipulability theory with something like the following:

> (**M**) *X* causes *Y* if and only if there are some possible interventions on *X* such that if they were to occur, *Y* would change.

Several additional comments about the notion of an intervention may be helpful. First, note that, as characterized above, the notion of an intervention makes no reference to distinctively human actions or to human abilities to manipulate—the notion is instead characterized in causal terms and in terms of facts about statistical independence. Thus some naturally occurring process P which bears the right causal and statistical relationships to *X* can count as intervention on *X* even if it does not involve human action (or at least deliberate human intervention) at any point. In this case, we have what is often described as a "natural experiment" – Mendelian randomization provides an example in the context of epidemiology (Davey Smith, Ebrahim, 2005). Manipulations carried out by human beings can of course qualify as interventions, if they have right sort of causal and statistical characteristics, but when they do so, this will be because of these characteristics and not because they are performed by humans. Introduction of the notion of an intervention helps to remove the anthropomorphism that otherwise infects manipulability accounts.

Second, it should be apparent that the notion of an intervention (like the more general notion of a manipulation) is a causal notion. Thus in appealing to this notion in order to explicate what it is for *X* to cause *Y*, we have given up on the project of providing a "reductive" account of causation. (Collingwood's account is also patently non-reductionist, since it makes use of such causally committed notions as "produce" and "prevent"). I will return below to the issue of how (as I think is the case) it is possible for a non-reductive account of causation to manage to be useful and illuminating.

Although (**M**) retains some of the basic commitments of Collingwood's treatment of "sense II" of cause, it avoids some difficulties faced by that account. First, pace Collingwood, whether (iii) *Y* changes (or is "produced" or prevented") under a human action that changes (or "produces or prevents") *X* is neither sufficient nor necessary for it to be true that (iv) *X* causes *Y*. (iii) is not sufficient for (iv) when the human action that changes *X* is confounded, as in example involving smoking above. With respect to the issue of whether (iii) is necessary for (iv), it seems to me to be central to how we think about causation (both in Collingwood's Sense II and whatever other "senses" of this notion may exist) that causal relationships can sometimes hold between factors in circumstances in which the manipulation of the cause is not within the power of any human being. As noted above, it seems entirely correct to say that the gravitational attraction of the moon is causally responsible for the behaviour of the tides, even though this causal agent is beyond the power of human beings to manipulate and this may well

always be the case. (**M**) accommodates this by reformulating the connection between causation and what happens under interventions as a *counterfactual* claim: for *X* to cause *Y*, it is not required that an intervention on *X* actually occur or be within human power, but merely that if some such an intervention *were* to occur, *Y* would change in some way. Even though (let us assume) human beings cannot carry out interventions that alter the position of the moon, it is none the less true that if such an intervention were to occur, this would be a way of affecting the tides.

Collingwood's own view of causal claims such as those about the effects of the position of the moon on the tides is that these involve a very different sense of "cause" than his sense II; he claims they employ a notion of cause (which he calls sense III) which is appropriate to "theoretical" as opposed to "natural" or "practical" science (p. 96) -- a notion according to which effects are invariably or unconditionally associated with their causes. Presumably, he thus would be undisturbed by the observation that his manipulation-based sense II of causation is unable to accommodate examples involving causes that humans are unable to manipulate. However, Collingwood's sharp separation between "practical" and "theoretical" science (and between sense II and sense III of "cause") seems unconvincing. The boundaries of what it is possible for human beings to manipulate obviously change with time (think of nuclear explosions, or space exploration) and it seems dubious that when a recognized causal relationship involving a cause that was previously unmanipulable (by humans) subsequently becomes manipulable, this involves a change in the "meaning" or sense of "cause" employed in that relationship. For example, when the once purely theoretical causal relationships (in Collingwood's sense) of Newtonian mechanics used to characterize the behaviour of objects at a great distance from the surface of the earth were used more recently as the basis for manipulations of space craft visiting the surface of Mars, it seems very natural to suppose that the very same causal relationships described theoretically by Newtonian mechanics are those we have now learned to exploit practically in manipulating the space craft. In other words, Newtonian mechanics seems to describe causal relationships in the sense of relationships that are potentially exploitable for manipulation and control (even if at a particular moment in time we lack the technology to exploit them) and are in this respect like Collingwood's sense II causal relationships rather than relationships that fall into some completely distinct category. So it makes sense to formulate a version of a manipulability theory that, like (**M**), accommodates this fact.

III.

Let me, by way of conclusion to this article, return to the issue of "circularity" raised above. I have already noted that a number of researchers have found it illuminating to connect causation and manipulation in broadly the way suggested by Collingwood. The puzzle this raises is how this connection can be fruitful, given that the notion of "manipulation" (and "intervention") seems to presuppose the very notion of causation that we are trying to explicate.

This is a difficult issue, deserving a more detailed treatment than I can give it here, but the following observations may be of some help. First, note that while it is true that notions like "manipulation" and "intervention" carry causal commitments, they need not carry the same causal commitments as the very causal relationships we seek to learn

about in performing experimental manipulations. That is, while to know that I have performed an intervention on *X* with respect to *Y*, I must have causal knowledge of various sorts (e.g., that the intervention is a cause of *X*, that it is independent of certain other causes of *Y* but not all such causes and so on), I do not have to know or to have already settled whether *X* causes *Y*, which is what I am trying to establish[3]. In other words, I can use some causal information (that an intervention on *X* with respect to *Y* has occurred) to establish whether some new, different causal claim (that *X* causes *Y*) is true. Indeed if this were not true, it is hard to see how we could ever learn about causal relationships by performing experiments. Thus, at least in this sort of case, no vicious circularity is present. Instead we have an illustration of the general principle that reliable inference to causal conclusions always requires additional background causal assumptions, in combination with other sorts of information that may be non-causal in character—e.g. correlational information. This principle is pithily captured in Cartwright's (1989) formulation: "no causes in, no causes out" (39).

Although this consideration may help to allay worries about unilluminating circularity in contexts in which an experimental manipulation is actually performed, it raises, as we noted above, a related question: How can a manipulability/interventionist conception of causation be useful or illuminating in situations in which an experimental manipulation is not actually performed—that is, in contexts in which one attempts to learn about causal relationships from non-experimental or "observational" data. (This of course is the typical context for causal inference in epidemiology.) The short answer to this question, articulated at length by a number of writers (including Pearl, 2000, Rubin, 1974), is that an interventionist account of causation can function fruitfully in such a contexts by specifying clearly what one is aiming at in causal inference and what is required to achieve this goal: in non-experimental contexts one tries to infer what would happen to some candidate effect if (contrary to actual fact) one were to perform an experiment in which an intervention on the cause occurs. In other words, one's goal is to infer the outcome of a hypothetical experiment without doing the experiment. (This counterfactual formulation is incorporated into **M**). This suggests, among other things, a standard for evaluating proposed causal inferences: an inference to a causal claim will be reliable to the extent that the inference warrants conclusions about what would happen if the hypothetical experiment associated with the experiment were to be performed. Thus, given a candidate causal inference, the question we should ask is: what sort of data is required and what additional assumptions must hold for the inference to establish the claimed result about the hypothetical experiment associated with the causal claim? (As noted above, such additional assumptions, carrying causal commitments, will *always* be necessary.) As an illustration, consider an inference from observational data that makes use of an *instrumental variable*: suppose that the problem we face is to estimate *b* in a context in which the data generating process is represented by $Y=bX+U$, X and Y are our candidates for cause and effect variables, respectively, and U is an error term that is

[3] This is somewhat obscured in Collingwood's formulation, since he speaks of "producing or preventing" the effect by manipulating the cause, which might seem to suggest that one has already established whether there is a causal relationship between cause and effect. A better formulation would require only that the candidate cause and effect are *correlated* under some interventions on the cause, as **M** does.

correlated with *X*. As is well-known, we can do this (despite the correlation between *X* and *U)* if we can find an instrumental variable *Z* which (i) is associated with *X* and (ii) is independent of *U* and (iii) is independent of *Y* given *X* and *U* . When these conditions are satisfied, it is widely accepted that the use of the instrumental variable Z allows us to estimate *b* and leads to reliable causal inferences. (See, e.g., Greenland, 2000, Angrist and Pischke, 2009). Thinking of causal claims within an interventionist/manipulationist framework makes it transparent why this is the case: under these conditions, *Z* functions in a way that is equivalent to an intervention on *X* with respect to *Y*, so that any variation in *Z* which is associated with *Y* must be due to the causal influence of *X* on *Y,* rather than being due to some other source, such as a confounding variable. Many other causal inference procedures can be evaluated in a similar way, by asking whether they provide reliable information about the results of a hypothetical experiment.

More generally, conceiving of causal claims as interventionists do, as having to do with the outcomes of hypothetical experiments, can play an important role in clarifying just what is meant by such claims and the evidence relevant to assessing them, even in contexts in which experiments cannot be performed. For example, the interverventionist framework forces us to specify candidate causal variables in such a way that they are possible targets of interventions, even if the interventions are not actually performed. In particular, this means that the "unit" or entity intervened on must be such that it is characterized by a variable that can assume at least two possible values, and that it must be possible to intervene to change one of these values to the other, with a corresponding change in the value of the effect. In other words, the cause variable must be a "treatment" variable and the effect variable should describe possible responses to that treatment. Not all candidates for causal variables will meet this condition. For example, instead of thinking of gender as a "cause" of differences in breast cancer rates among men and women, it is more perspicuous to think of the relevant causal variable as having to do with differences in the level of endogenous estrogens. (cf. Joffe et al., 2012.) The latter, unlike gender, is a variable which is a well-defined target of manipulation and hence causal claims formulated in terms of this variable make it clearer what the associated hypothetical experiment is and what is claimed about the outcome of this manipulation.

For all of these reasons, manipulability accounts of causation of the sort sketched by Collingwood remain a fruitful framework for thinking about causation.

References

Cartwright, N. *Nature's Capacities and their Measurement.* 1989. Oxford: Oxford University Press.

Collingwood, R.G. On the So-Called Idea of Causation. *Proceedings of the Aristotelian Society* 1937; **38**:85-112.

Davey Smith, G. , Ebrahim, S. What Can Mendelian Randomization Tell Us About Modifiable Behavioural and Environmental Exposures? *BMJ*. 2005; **330**:1076-79.

Gasking, D. Causation and Recipes. *Mind* 1955; **64**: 479-87.

Greenland, S. An Introduction to Instrumental Variables for Epidemiologists. 2000. *International Journal of Epidemiology* **29**; 722-29.

Hausman, D. Causal Asymmetries.1998. Cambridge: Cambridge University Press.

Joffe, M., Gambhir, M. Chadeau- Hyam, M. Vineis, P. Causal Diagrams in Systems Epidemiology. 2012. *Emerging Themes in Epidemiology* **9**:1. http:/www.ete-online.com/content/9/1/1.

Price, H. Agency and Probabilistic Causality. *British Journal for the Philosophy of Science*.1991; **42**:157-76.

Pearl, J. Causality: Models, Reasoning, and Inference. 2000. Cambridge: Cambridge University Press.

Rubin, D. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology.* 1974; **66**: 688-701.

Spirtes, Glmour and Scheines. Causation, Prediction, and Search. 2000. Cambridge, MA: MIT Press.

Von Wright, G. Explanation and Understanding 1971. Ithaca, NY: Cornell University Press.

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