Interdefining Causation and Intervention

Michael BAUMGARTNER[†]

ABSTRACT

Non-reductive interventionist theories of causation and methodologies of causal reasoning embedded in that theoretical framework have become increasingly popular in recent years. This paper argues that one variant of an interventionist account of causation, *viz.* the one presented, for example, in Woodward (2003), is unsuited as theoretical fundament of interventionist methodologies of causal reasoning, because it renders corresponding methodologies incapable of uncovering a causal structure in a finite number of steps. This finding runs counter to Woodward's own assessment and to other recent studies which presume that Woodward's version of interventionism is effectively applicable to uncover causal structures, e.g. Campbell (2007) or Shapiro and Sober (2007).

I Introduction

Interventionist theories of causation are typically subdivided into two categories: *reductive* and *non-reductive* accounts.[1] Reductive theories, as advanced by Collingwood (1940), Gasking (1955), von Wright (1971) or Menzies and Price (1993), reduce the notion of causation to an allegedly non-causal notion of intervention or manipulation, while according to non-reductive accounts, as can be found in Spirtes et al. (2000), Pearl (2000) or Woodward (2003), such a reduction is not possible. Reductive theories have been broadly criticized for a number of reasons, e.g. for being unacceptably anthropocentric or circular (cf. Mackie 1976 or Hausman 1986). In consequence, they have only played a marginal role in the causation literature of the second half of the 20th century. Non-reductive variants of interventionist theories, however, have become increasingly popular in recent years. Especially the literature focussing on matters of causal reasoning in various disciplines is more and more drawing on ideas developed in this non-reductive interventionist framework.

Non-reductive theories come in two groups. The first group is constituted by accounts that analyze the notion of intervention in terms of causation which, in turn, is introduced as a primitive or unanalyzed concept (cf. e.g. Spirtes et al. 2000,

[†]Department of Philosophy, University of Bern, Laenggassstr. 49a, 3012 Bern, Switzerland; Email: baumgartner@philo.unibe.ch

[1]Cf. e.g. Hausman and Woodward (1999), pp. 533-534.

sect. 3.7.2 and 7.5, or Pearl 2000) – for easy reference, call the theories in this group *p-theories*. The second group comprises theories that maintain a tight conceptual interdependence between the notions of causation and intervention by, very broadly, spelling one of the two notions out in terms of the other and vice versa (cf. Woodward 1997, Hausman and Woodward 1999, Hausman 1998, sect. 5.3* and 7.1* , Woodward 2003, Woodward 2007a). That means the theories in this second group conceive of causation and intervention as two *interdefined* concepts – call them *i-theories* for short. As suggested in Woodward (2003, 104–107), the particular conceptual interdependence of causation and intervention advocated by i-theories is not viciously circular. Moreover, we shall see in section III below that the conceptual core of i-theories has some very specific, indeed, rather strong implications, and thus is far from being empty. That is, even though i-theories interdefine causation and intervention, they can be argued to be informative.

Merely shedding light on the conceptual interdependence of causation and intervention, however, is not the main aim of i-theories. Rather, Woodward (2008, 194) claims that their "primary focus is *methodological*". More specifically, he intends his i-theory to illuminate "how we think about, learn about, and reason with various causal notions" (Woodward 2008, 194). Woodward (2003, chs. 1 and 3) takes the i-theoretical framework to provide the means to experimentally uncover causal structures that involve variables whose values are actually manipulable or to test corresponding causal claims. He holds that i-theories permit to distinguish experimentally, at least in some cases, between causal structures that are indistinguishable based, for example, on statistical data alone. Several authors have followed Woodward in judging that i-theories are effectively applicable in experimental contexts. Shapiro and Sober (2007), for instance, hold that Woodward's version of interventionism provides a means to experimentally identify micro-effects of macro-causes. Or Campbell (2007) accounts for empirical investigation into psychological causation on the basis of Woodward's i-theory. Such studies presume that the interdefined conceptual fundament of i-theories is fruitfully applicable to the discovery of causal structures. According to this presumption, i-theories not only clarify the conceptual interdependence of causation and intervention but can also be resorted to when it comes to grounding a methodology of causal reasoning. That is, the fact that i-theories interdefine causation and intervention in an informative way and, thus, can be said to be *conceptually unproblematic* is taken to imply that applying i-theories to uncover causal structures is methodologically or *epistemically unproblematic* as well.

The paper at hand takes issue with this claim. Its main goal is to show that the interdefined conceptual fundament of i-theories, notwithstanding the fact that it is conceptually informative, gives rise to a severe *epistemic* problem when i-theories are resorted to in the course of causal discovery. For applying the conceptual core of i-theories to concrete causal processes in experimental contexts triggers infinite regresses that render it impossible, in principle, to determine of even two concrete variables whether they are causally connected or not in a finite number of steps. Hence, this paper is going to argue that i-theories are not effectively applicable to

solve problems of causal discovery. As we shall see, this finding does not generally call into question the power and effectiveness of interventionist methodologies of causal reasoning as, for example, presented in Spirtes et al. (2000) or Pearl (2000), which often clearly outperform alternative methodologies. Rather, the epistemic regresses triggered by an application of i-theories in experimental contexts demonstrate that efficient interventionist methodologies cannot and must not be based on the interdefined conceptual core of i-theories. At least two alternative conceptual foundations of interventionist methodologies remain possible: Either causation is introduced as a primitive notion, as done by p-theories, or causation is spelled out in non-interventionist terms, say in probabilistic or regularity theoretic terms. Intervening on causal structures is one of the most effective ways to uncover these structures. For this very reason, however, the itself causally loaded notion of intervention is unsuited as definiens of causation in the context of causal discovery.

In section II, the core of i-theoretical interventionism as presented in Woodward (2003) is briefly reviewed. Section III then shows that even though the conceptual fundament of Woodward's variant of interventionism is illuminating, its application in experimental contexts triggers infinite regresses which give rise to a severe epistemic problem if methodologies of causal reasoning are grounded on an i-theoretical foundation. In section IV, two conceivable strategies to stop the epistemic regresses are discussed and demonstrated to fail. The paper concludes that there cannot exist an effective interventionist methodology that uncovers causal dependencies as defined by i-theories.

II Interdefined Concepts

The by far most thorough and elaborate presentation of an i-theory can be found in Woodward (2003).[2] Woodward's theory turns on two core definitions. First, he defines the notions of a direct and of a contributing cause (p. 59):

(M) A necessary and sufficient condition for X to be a (type-level) *direct cause* of Y with respect to a variable set $\mathbf{V}$ is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in $\mathbf{V}$ (by interventions)[3]. A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set $\mathbf{V}$ is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; (...) and that (ii) there be some intervention on X that will change Y when all other variables in $\mathbf{V}$ that are not on this path are fixed at some value. (...)

[2]Consequently, Woodward (2003) constitutes the central point of reference for all attempts to uncover causal structures based on methodologies rooted in i-theories.

[3]This addition is not explicitly mentioned in (M). It is contained, however, in the separate definition of a direct cause (DC) given on p. 55.

Against this background, a variable X is a cause of Y iff X is either a direct or a contributing cause of Y. Second, Woodward (2003, 98) defines the notion of an intervention variable:

(IV) I is an intervention variable for X with respect to Y iff

1. I causes X;
2. I acts as a switch for all other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I;
3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the $I - X - Y$ connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X;
4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.

Finally, relative to the notion of an intervention variable, an (actual) *intervention* can be straightforwardly understood in terms of an intervention variable I for X with respect to Y taking on some value z_i such that $I = z_i$ causes X to take on some determinate value z_j (p. 98).

In several passages, Woodward (2003, 55, 60–61, 98) explicitly refers to (M) and (IV) as *definitions* of the pertaining notions and he claims that (M) and (IV) provide *truth conditions* for causal statements as "X causes Y".[4] In regard to defining causation in terms of intervention, and vice versa, he writes (p. 61):

> In other words, once we fix our representational repertory (i.e., once we choose a set of variables to represent the quantities whose causal relationships we are interested in assessing), then two theories will make different claims about causal relationships among these variables if and only if they make different claims about what will happen under some combination of interventions. Putting this in the form of a slogan, we can say that manipulability accounts are committed to the following: *No causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference*.

Several things need to be noted about a theory of causation that turns on (M) and (IV). First, it is clearly non-reductive as it does not spell out causation in non-causal terms. Second, an i-theory based on (M) and (IV) differs from traditional

[4]Note that in a recent response to Strevens (2007), Woodward (2008, 195) insists that his use of the term *definition* does not carry metaphysical implications. More concretely, by interdefining causation and intervention, Woodward does not intend to claim that these concepts metaphysically depend on each other.

reductive interventionist accounts in not involving the notion of human action. (IV) yields a notion of an intervention variable that is thoroughly non-anthropocentric. An intervention variable is solely defined in terms of its causal (and statistical) relations to the other variables in a given structure. And third, it is a variant of a counterfactual analysis of causation because the notion of a *possible* intervention contained in (M), according to Woodward (2003, 70–71), "should be interpreted to mean that there is some intervention on X such that *if it were possible to intervene to manipulate X repeatedly in that way*, Y (or the probability of Y) would change in some reproducible or repeatable way".

III Regresses

Obviously, causation and intervention are interdefined by (M) and (IV). That, however, is not considered to be problematic by Woodward. He demonstrates that the particular conceptual interdependence of causation and intervention induced by (M) and (IV) is far from being vacuous and, moreover, he maintains that it is not viciously circular (Woodward 2003, 104–105):

> The causal information required to characterize the notion of intervention on X with respect to Y is information about the causal relationship between the intervention variable I and X, information about whether there are other causes of Y that are correlated with I, information about whether there is a causal route from I to Y that does not go through X and so on, *but not information about the presence or absence of a causal relationship between X and Y*.

Woodward clearly is right that (M) and (IV) do not analyze "X causes Y" by drawing on "X causes Y" itself, but rather by drawing on e.g. "I causes X" or "I does not directly cause Y" which are different causal (in-)dependencies. Still, the question whether that suffices to render the conceptual core of Woodward's theory non-circular has provoked some controversies in the literature. De Regt (2004) and, most of all, Strevens (2007, 2008) argue that (M) and (IV) give rise to circularities, notwithstanding the fact that in neither of the two definitions "X causes Y" appears on both sides of the biconditional. Woodward (2008), in return, emphatically insists on the non-circularity of the conceptual foundation of his theory – and the majority of writers have followed him in this regard. I shall not enter into this debate here. All that matters for our current purposes is the uncontroversial fact that the interdependence of causation and intervention as expressed in (M) and (IV) does not result in an empty theory. An i-theory centered around (M) and (IV) has some very specific implications. For instance, two variables, neither of which can be intervened upon, are not causally related according to such a theory. To illustrate, take a variable representing a supervenient property as exemplified by a mental phenomenon M. Suppose, we want to determine whether M causes some physical effect, say an action A. If M is seen to represent a supervenient property, it cannot be manipulated without at the same time changing its physical

supervenience base P. The latter, however, is supposedly located on a causal path leading to A that does not contain M. Since intervening on M is correlated with changes in P which cannot be held fixed by interventions while manipulating M, it is impossible that there exists an intervention variable I for M with respect to A, for condition (IV.4) cannot be satisfied. Moreover, if manipulations of M are not only seen to be correlated with changes in P but also to cause these changes (independently of M), condition (IV.3) cannot be met either. The fact that there is no possible intervention on M with respect to A, according to (M), implies that M does not cause A. Thus, (M) and (IV) imply that, if mental phenomena are seen to exemplify supervenient properties, there is no mental-to-physical causation.[5] Put differently, if somebody thinks that there exists mental-to-physical causation and that causation is to be understood in terms of (M) and (IV), he or she cannot conceive of mental phenomena as instantiations of supervenient properties. Or, to take a different consequence of (M) and (IV) mentioned in Woodward (2007b, 22): (M) and (IV) are "inconsistent with many other claims made about causation, for example, claims that causal relationships require a spatiotemporally connecting causal process". Thus, notwithstanding the direct conceptual interdependence of (M) and (IV), Woodward's i-theory has certain rather strong implications.

Whatever one's attitude towards these implications may be, the way the notions of causation and intervention are interdefined in Woodward's i-theory is not circular to the effect that the theory would be rendered uninformative. It can be justly argued that (M) and (IV) illuminate the conceptual interdependence of these two causally entrenched notions. Woodward (2003, chs. 1 and 3), however, does not content himself with clarifying the conceptual interdependence of causation and intervention. Rather, he takes his version of interventionism to be effectively applicable, at least in some experimental contexts, to determining whether a causal relationship exists between two variables or to testing truth-values of causal claims. Woodward assumes that because the interdefinition of causation and intervention can be argued to be *conceptually unproblematic* the application of the resulting version of interventionism to (experimentally) uncovering causal structures is *epistemically unproblematic* as well. The remainder of this paper shall cast doubts on the accuracy of this presumption which, as shown in section I, is shared by a number of authors.

Even though the conceptual foundation laid out in (M) and (IV) is not vacuous, the fact that (M) and (IV) interdefine causation and intervention raises a pressing *epistemic* problem when it comes to causal reasoning on the basis of (M) and (IV). Experimentally uncovering causal structures by interventionist means essentially draws on systematic manipulations of these structures (or parts thereof) with the use of intervention variables.[6] In this vein, it often becomes possible to disambiguate causal inferences that would remain ambiguous were it not for the avail-

[5] In Baumgartner (forthcoming), I present this interventionist exclusion argument in all detail and show that it rests on considerably weaker premises than classical exclusion arguments.

[6] Cf. e.g. Spirtes et al. (2000) or Pearl (2000).

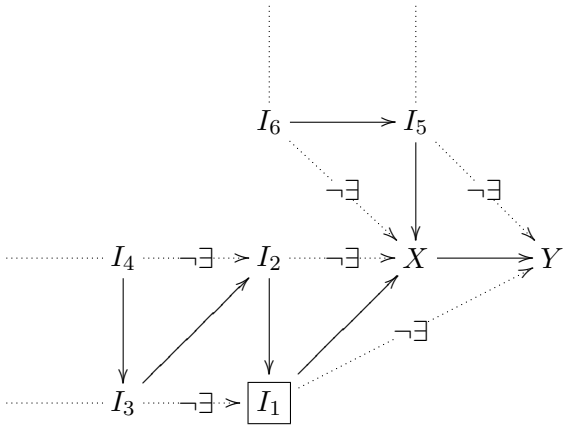

Figure 1: Two infinite regresses induced by interdefining causation and intervention.

ability of intervention variables. That means fruitfully applying (M) and (IV) to problems of causal discovery, first and foremost presupposes that it is possible to identify certain variables in an investigated structure as intervention variables.

Hence let us see how intervention variables could actually be identified within the i-theoretical framework. Suppose we want to determine whether a variable I_1 is an intervention variable for X with respect to Y. Condition (IV.1) stipulates that a necessary condition for that to be the case is I_1 being a cause of X. According to (M), a necessary condition for I_1 to be a cause of X is that there be a possible intervention, call it I_2, on I_1 with respect to X. This, in turn, requires I_2 to be a cause of I_1, which again presupposes that there is a possible intervention I_3 on I_2 with respect to I_1 which calls for a further possible intervention I_4 on I_3 with respect to I_2, and so on. Condition (IV.3) amounts to another necessary condition for I_1 to be an intervention variable for X with respect to Y: There must not be a causal path connecting I_1 and Y that does not go through X. In order to determine whether I_1 satisfies that condition, firstly the possible $I_1 - X - Y$ connection must be suppressed (or 'broken') by intervening on X by means of a further intervention variable I_5 and, secondly it must be established that there is no possible intervention on I_1 such that Y covaries with I_1.[7] Of course, according to (M), I_5 being an intervention variable for X with respect to Y requires there to be another possible intervention I_6 on I_5 with respect to X, and so on. In sum, validating that conditions (IV.1) and (IV.3) are satisfied, given that causation is defined in terms of (M), triggers at least two infinite regresses.[8] For illustrative purposes these regresses are schematically graphed in figure 1. Since the notion of causation also is of crucial importance in (IV.2) and (IV.4) similar regresses are initiated when it

[7] For further details on testing the satisfaction (IV) cf. Woodward (2003, 99–111).

[8] For a criticism of Woodward (2003) along similar lines cf. Strevens (2007).

comes to determining whether a specific variable I_1 satisfies these conditions. For easy reference later on, call these regresses *identification regresses*.

So even though Woodward can justifiably argue that interdefining causation and intervention as done in (M) and (IV) is conceptually informative, that definitional interdependence renders it impossible to ever establish one single variable as an intervention variable in a finite number of steps. As long as one were only concerned with spelling out the conceptual relationship between causation and intervention, these identification regresses would not seem particularly problematic. However, as soon as causal structures are actually to be uncovered on the basis of (M) and (IV), it becomes of utmost importance that concrete variables be identifiable as intervention variables. (M) and (IV) can only be effectively applied to determining whether a causal relationship exists between two variables or to testing truth-values of causal claims, if the infinite regresses compromising the identification of intervention variables can somehow be stopped. In sum, whoever joins Woodward in taking (M) and (IV) to be profitably applicable to problems of causal discovery has to answer the question as to how to stop these regresses.

IV How to Stop the Regresses?

Although Woodward does not explicitly discuss identification regresses, he provides some indications as to how somebody professing an i-theory that turns on (M) and (IV) could answer the question raised in the previous section (cf. Woodward 2003, ch. 3). For instance, Woodward is very clear about the fact that his account does not reduce causal to non-causal information. Moreover, he takes such a reduction to be impossible in principle and, accordingly, his interventionist analysis, inter alia, to merely express that impossibility. In accordance with this claim, there does not exist a methodology of causal reasoning that infers causal information from purely non-causal information. All available methodologies either require prior causal knowledge or adopt causal assumptions about the causal background of investigated structures, about the structures themselves, or about how analyzed data reflect underlying structures.

These considerations suggest two feasible strategies to stop identification regresses: An interventionist methodology of causal reasoning embedded in (M) and (IV) gets off the ground only if either (i) it is possible to draw on prior causal knowledge about a structure under investigation that establishes that certain variables have the interventionist properties or if (ii) pertaining variables are simply assumed (without being known) to comply with (IV) relative to (M).[9] For epistemic reasons, *knowing* that a specific variable has the interventionist properties is certainly preferable to merely *assuming* that to be the case. However, such prior causal knowledge will not generally be available. In that case, (ii) is the only vi-

[9] Strategy (ii) reflects what Glymour (2004) has dubbed the Euclidean approach to causal discovery, whose main interest consists in developing axiomatic systems of causal reasoning. For more details cf. section IV.2 below.

able strategy to stop the identification regresses. The two strategies are logically independent. Thus, it may turn out that they are both successful or that none of them is or that one of them is while the other is not. The following two subsections investigate the prospects of stopping identification regresses by either drawing on prior causal knowledge or by assuming variables to be of the interventionist type, i.e. by employing strategies (i) and (ii), respectively.

IV.1 Prior Knowledge

If a variable I contained in a causal structure $\mathcal{S}$ is in fact known to satisfy (IV) relative to the analysis of causation given in (M) prior to investigating $\mathcal{S}$, there obviously is no need to apply (M) in order to re-establish that I complies with (IV). In consequence, no regress is set off immediately. Accordingly, interventionists employing strategy (i) typically advocate their approach by arguing that there indeed exist concrete variables that are known to satisfy (IV). A typical vindication of strategy (i) along these lines can be found in Woodward's (2008, 203-204) response to Strevens (2007), where Woodward maintains that certain variables involved in randomized experiments undoubtedly are intervention variables.[10] Suppose we want to find out whether treatment with a specific drug (T) is a cause of recovery from a particular disease (R). In order to answer that question, subjects that suffer from the disease are randomly assigned to treatment and control groups, say, by tossing a coin (C). Everybody will agree that C is an intervention variable for T with respect to R. C determines whether somebody is assigned to treatment or control group, i.e. C's value is sufficient for the value of T, C breaks all other arrows into T, C does not directly cause R, and C is independent of other causes of R. We have enough prior causal knowledge to be reasonably confident that the triple $\langle C, T, R\rangle$ satisfies (IV). Thus, there are variables which we can conclusively identify as intervention variables and, therefore, the identification regresses do not seem to be insurmountable.

The fact that we seem to have enough prior causal knowledge to positively identify coin tossings as intervention variables, notwithstanding the identification regresses exhibited in the previous section, of course, raises the question as to where that knowledge comes from and what evidence it is based on. What is the i-theoretical rationale for knowing that C is an intervention variable for T with respect to R? The interventionist framework provides two conceivable justificatory sources for such knowledge: either this prior knowledge is justified *by direct application of (M) and (IV)* to the triple $\langle C, T, R\rangle$ or – if that is not possible – it is shown that there exist *suitable heuristics* that ground such knowledge by ascertaining that $\langle C, T, R\rangle$ complies with (M) and (IV) without direct application of (M) and (IV). Let us take a closer look at these two possibilities to ground the causal knowledge required to stop identification regresses.

[10] Similarly, Woodward (2003, 94–98).

What would establishing or justifying the knowledge that C is an intervention variable for T with respect to R by application of (M) and (IV) amount to? According to (IV), knowing that C is an intervention variable for T with respect to R, among other things, presupposes that C is known to be a cause of T. This latter knowledge, according to (M), requires that one knows that there is a possible intervention I_1 on C with respect to T which, according to (IV), presupposes that one knows that I_1 causes C, which again requires that one knows that there is a possible intervention I_2 on I_1 with respect to C – and so forth. That is, answering the question as to what are the i-theoretical conditions for knowing that a certain variable is an intervention variable based on (M) and (IV) triggers regresses all anew. Attempting to justify prior interventionist knowledge by directly applying (M) and (IV) to concrete variables does not stop the regresses but merely dislocates them. Or it might be said that strategy (i) just initiates different regresses: applying (M) and (IV) to a specific variable in order to determine whether it is an intervention variable sets off identification regresses, whereas answering the question as to what warrants prior interventionist knowledge by drawing on (M) and (IV) sets off what I shall subsequently call *justification* regresses. It is not only impossible to identify intervention variables on the basis of (M) and (IV), it is also impossible to justify prior interventionist knowledge by applying (M) and (IV).

As indicated above, however, in order to determine whether a specific entity satisfies a given definition it is often not necessary to apply the definition itself, rather, heuristics will do. If I want to know whether a yellow ring is made of gold, I do not necessarily have to conduct a chemical analysis. Many suitable heuristics are available. The price of the ring will be an indication, the reputation of the store in which it is sold or of the person that sells it or some engraving on it might enable me to decide the matter. The interventionist could, thus, insist that determining whether variables are intervention variables in the sense of (IV) relative to (M) could, analogously, be delegated to suitable heuristics. Hence, how could the satisfaction of (M) and (IV) be assessed without applying (M) and (IV) themselves?

Causation, as is well known, does not necessarily have to be spelled out in interventionist terms. There are many alternative theories available in the literature, and relative to some of them coin tossings can indeed be straightforwardly identified as intervention variables for treatment with respect to recovery. Take for instance an elementary probabilistic analysis as professed by Suppes (1970). Given a suitable probability distribution over C, T and R, such a theoretical framework identifies C as direct cause of T if C is positively correlated with T, C occurs prior to T, and C is not screened off from T by some third variable in the structure. Furthermore, if C is screened off from R by T, C can be said not to directly cause R. Finally, if pertaining probabilistic data can be shown not to feature any other (probabilistically defined) causes of R that are correlated with C, it follows that the triple $\langle C, T, R\rangle$ satisfies all conditions given in (IV) and, thus, that C is an intervention variable for T with respect to R. Of course, such a probabilistic analysis in the vein of Suppes (1970) has long been shown not to adequately capture

all causal dependencies, as e.g. causes that lower the probabilities of their effects. That, however, is not at issue here. Rather, what is important for the present context is that Suppes' theory, irrespective of whether it successfully accounts for *all* kinds of causal dependencies, very straightforwardly allows for assessing the satisfaction of (IV) by *coin tossings* in randomized experiments. Modern and more sophisticated probabilistic analyses as e.g. professed in Kvart (1997, 2001, 2004) would also allow to establish C as intervention variable for T with respect to R in a finite number of steps – even though such an assessment would involve more complications. Besides, in case of other sorts of examples as e.g. mechanical ones, variables could also be demonstrated to satisfy (IV) in a finite number of steps by employing a transference or process theoretic account, or in deterministic cases a regularity theoretic analysis might do the job.[11]

Such alternative theories of causation, of course, come with their own and, most of all, non-interventionist definitions of causation. That is, establishing that the triple $\langle C, T, R\rangle$ satisfies (IV) by, say, probabilistic means in the vein of Suppes does not amount to showing that $\langle C, T, R\rangle$ satisfies (IV) relative to the notion of causation defined in (M) – it only amounts to showing that $\langle C, T, R\rangle$ complies with (IV) relative to the definition of causation given in Suppes (1970). Nonetheless, the i-theorist could claim that C being a direct cause of T and no direct cause of R subject to a probabilistic analysis can be seen as a heuristic measure of C being a direct cause of T and no direct cause of R subject to (M). Hence, in light of the impossibility to justify interventionist causal knowledge by direct application of (M) and (IV) the i-theorist could advance *non-interventionist* accounts as heuristics for assessing the satisfaction of (M) and (IV) without direct application (M) and (IV).

Clearly, heuristics often render it unnecessary to explicitly apply definitions. Yet, there is an important difference between considering a ring's price as a heuristic measure of its chemical structure and, for example, using probabilistic procedures as heuristics for uncovering causal dependencies in the sense given by (M). Whenever the question whether a ring is made of gold is answered by looking at its price, it would, at least in principle, be possible to conduct a chemical analysis and, thus, to explicitly apply the definition of gold. That is fundamentally different in case of causal dependencies defined along the lines of (M). The only way to determine whether C is an intervention variable for T with respect to R in a finite number of steps – as the above considerations suggest – is to apply some non-interventionist account of causation, i.e. to apply what the interventionist would like to see as a mere heuristic for causation. This, in turn, casts serious doubts on the heuristic character of non-interventionist approaches to identifying intervention variables. For in order to establish a certain non-definitional property of an entity of type t as a heuristic measure for the identification of entities of type t, it must be shown that the non-definitional property indeed *coincides* with the

[11] For transference and process theoretic analyses cf. e.g. Salmon (1998) or Dowe (2000), for a regularity theoretic account cf. Baumgartner (2008a).

definitional properties of entities of type t. That is only possible if at least some entities of type t can actually be identified by explicitly applying t's definition. That is, heuristics for t can only be validated if the definition of entities of type t is applicable in a finite number of steps, at least in principle. While that condition is certainly satisfied in case of gold and its price, every application of the interventionist definition of causation in the course of identifying intervention variables triggers infinite regresses. There is no way to identify intervention variables or causal dependencies by applying (M) and (IV) in a finite number of steps in even one single case. In view of this lack of a single positive application of (M) and (IV), non-interventionist accounts cannot be given the status of heuristics for assessing the satisfaction of (M) and (IV). Instead, they provide *self-contained analyses* of causation that are *independent* of the notion of intervention. There cannot exist a heuristic for (IV)-defined intervention relative to (M)-defined causation because there does not exist a single variable that can actually be shown to satisfy both (IV) and (M).

Given that assessing whether the triple $\langle C, T, R\rangle$ satisfies (M) and (IV) hinges on an infinite array of presuppositions, it is impossible to ever be reasonably confident that even a single one of these presuppositions is actually satisfied in a particular case. Yet notwithstanding those i-theoretical regresses, we are as certain as one can possibly be in empirical matters that coin tossing indeed is a form of intervening on whether patients are assigned to treatment or control group. This widespread certainty should appear completely mysterious to the i-theorist, for her analysis does not provide any rationale for such interventionist knowledge. In fact, however, our unshakable conviction that C is an intervention variable for T with respect to R is neither mysterious nor ill-founded, nor does it prove that i-theoretical regresses can be stopped. Rather, it simply shows that *de facto* we do not understand causation in terms of (M). Whoever is convinced that coin tossing is an intervention variable for treatment with respect to recovery *cannot* and, as a matter of fact, *does not* understand causation in terms of (M), but rather in terms of some non-interventionist account. As soon as the interdefinition of causation and intervention is removed, all regresses – and with them, all problems we have been dealing with so far – disappear. Coin tossing can be straightforwardly established as intervention variable based on virtually any non-interventionist account of causation.

It might be argued, at this point, that these objections to building a methodology of causal reasoning on (M) and (IV) implicitly presuppose the ideal of some kind of foundationalist epistemology which cannot be provided in principle and which, hence, leads the way into mere skepticism. In consequence, the interventionist could reply that even the identification of gold does not terminate with self-evident givens or truths. Rather, a gold-identifying chemical analysis draws on certain causal characteristics of gold as, for instance, that exposing gold to nitric acid – contrary to exposing mere base metals to nitric acid – neither causes changes in color nor dissolution. Depending on theoretical preferences, these causal characteristics presuppose, say, probabilistic independencies (plus additional non-causal

empirical information as e.g. temporal orderings) to the effect that exposure of gold to nitric acid does not raise the probability of gold changing its color or dissolving. These probabilistic independencies, again depending on theoretical preferences, can be claimed to rely on e.g. frequencies such that exposure of gold to nitric acid is not correlated with gold changing its color or dissolving; and so forth. It might thus be held that even the identification of gold triggers a regress and, accordingly, is by no means better off than the identification of intervention variables along the lines of (M) and (IV).

The above criticism of the i-theoretical framework, however, in no way presumes the non-attainable ideal of a foundationalist grounding of all human knowledge. Theorizing about the world never starts from scratch, but always takes some conceptual frame as given and unquestioned. It is exactly this inevitable grounding of all knowledge in conceptual primitives that guarantees that the regress prompted by the application of a chemical definition of gold is *not infinite*, but terminates as soon as some conceptual level is reached that is considered to be primitive by whoever happens to apply that definition. To a chemist the gold-identifying chemical analysis, most likely, is beyond doubt; a philosopher professing a p-theory of causation will be satisfied if the analyzed ring has been shown to have the causal characteristics of gold; somebody opting for a probabilistic account of causation will further want to see these causal characteristics reduced to probabilistic dependencies; and, finally, somebody advancing a frequentist interpretation of probabilities will require a reduction of probabilistic dependencies to a suitable frequency distribution. What is of crucial importance here is that all levels of this conceptual reduction of the notion of gold are *independent* of their preceding levels. Whereas the application of the definition of gold to a ring induces a progression from one conceptual level to a subsequent independent level and stops when some primitive level is reached, the application of (M) in order to identify intervention variables as defined by (IV) induces an *infinite oscillation* between two notions, none of which is taken to be primitive by a corresponding i-theory. The identification and justification regresses triggered by (M) and (IV) never reach a primitive conceptual level and are, thus, infinite.

All in all, the prospects of blocking the identification regresses by drawing on prior causal knowledge look dim. Rather than showing how these regresses can be blocked, the discussion of this section suggests that prior knowledge about the interventionist properties of a specific variable can only be established if the interdependence of causation and intervention as defined by (M) and (IV) is broken. As long as causation and intervention remain interdefined, justifying prior interventionist knowledge triggers analogous regresses as identifying intervention variables. All of these problems disappear immediately if causation is no longer defined in terms of (M) but in any way that is independent of (IV) – or if it is treated as primitive.

IV.2 Causal Assumptions

Let us now turn to strategy (ii) to answer the question raised in the previous section. According to that strategy, the identification regresses launched by applying (M) and (IV) are stopped by simply assuming certain variables to function as (M)- and (IV)-defined intervention variables within an investigated structure. There does not currently exist a methodology of causal reasoning that infers causal structures from purely non-causal empirical data. Some causal assumptions are presupposed by any available methodology. For instance, Boolean methodologies assume the causal backgrounds of analyzed data to be homogenous, or methodologies that analyze causal structures in terms of Bayesian networks assume causal structures and the probability distributions they generate to satisfy the *causal Markov* and *faithfulness* assumptions.[12] Many other types of causal assumptions can be found in the literature. What is important is that Cartwright's famous dictum "No causes in, no causes out" is often seen as being a sort of truism of causal data analysis. Thus, the interventionist might argue that simply assuming (without actually knowing) that a certain variable functions as an (M)- and (IV)-defined intervention variable in an investigated structure is as good a causal assumption as any other causal assumption entering causal reasoning.

This strategy receives additional support from a distinction between two different sorts of theories of causation that has been introduced by Glymour (2004, 779):

> Philosophical theories come chiefly in two flavors, Socratic and Euclidean. Socratic philosophical theories, whose paradigm is *The Meno*, advance an analysis (sometimes called an 'explication'), a set of purportedly necessary and sufficient conditions for some concept, giving its meaning; in justification they consider examples, putative counterexamples, alternative analyses, and relations to other concepts. Euclidean philosophical theories, whose paradigm is *The Elements*, advance assumptions, considerations taken to warrant them, and investigate the consequences of the assumptions. Socratic theories have the form of definitions. (...) Euclidian [sic!] theories have the form of formal or informal axiomatic systems and are often essentially mathematical (...)

If an interventionist theory of causation is conceived to be of the Euclidean type, its goal is not a conceptual analysis of causation or not even a conceptual clarification of how the notions of causation and intervention relate to each other, rather it can be seen to constitute the core of an axiomatic system of causal reasoning. Against such a background, when it comes to causal discovery by means of

[12] For a Boolean methodology cf. Baumgartner (2008b); for procedures uncovering causal Bayesian networks cf. Spirtes et al. (2000). The *causal Markov assumption* states that in a probability distribution $\mathcal{P}$ generated by a (acyclic) causal structure $\mathcal{S}$ a variable Z is independent of all its non-effects in $\mathcal{S}$ conditional on all of Z's direct causes, provided that no direct common causes of any two variables in $\mathcal{S}$ are left out of $\mathcal{P}$. According to the *faithfulness assumption*, there are no other conditional independence relations in $\mathcal{P}$ than the ones implied by the causal Markov assumption.

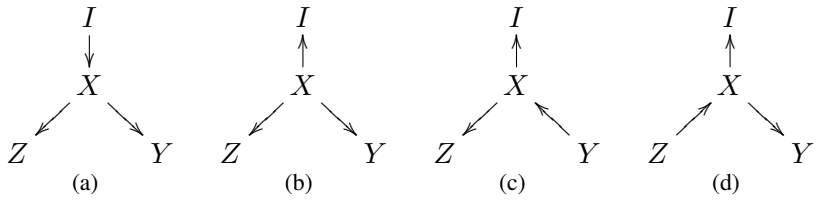

Figure 2: Four causal structures that all could generate the data represented in (1).

intervention variables, certain variables may, in a sense, axiomatically be assumed to be of the interventionist type in order to determine what can be learned about an investigated causal structure by means of these assumptions. Should it then turn out that such assumptions successfully disambiguate otherwise ambiguous causal inferences, the assumptions could be argued to be sufficiently warranted.

Such as to illustrate this axiomatic approach, suppose we are investigating the causal structure behind four variables I, X, Y, and Z. A study is conducted that presumedly yields the following independence relations as empirical data (cf. Eberhardt et al. 2006):[13]

$$Y \perp\!\!\!\perp Z \mid X; \quad I \perp\!\!\!\perp Z \mid X; \quad I \perp\!\!\!\perp Y \mid X \tag{1}$$

Thus, the structure behind I, X, Y, and Z is of probabilistic nature. The most sophisticated and efficient procedures to uncover causal structures from probabilistic data are embedded in a theoretical framework according to which causal structures are to be analyzed in terms of Bayesian networks (cf. e.g. Spirtes et al. 2000) – hence, all these discovery algorithms shall be referred to as *BN-algorithms* in the following. According to that framework, the empirical data collected in our hypothetical study could be generated by either of the four structures depicted in figure 2, because these four structurings of I, X, Y, and Z yield exactly the independence relations given in (1). Correspondingly, given input data as in (1) the output of BN-algorithms (roughly) corresponds to the set constituted by these four structures. That is, the empirical data recorded in (1) significantly underdetermines causal reasoning. It is ambiguous which of the graphs in figure 2 adequately represents the causal structure behind I, X, Y, and Z.

This ambiguity cannot be resolved solely based on the probabilistic data recorded in (1). The interventionist framework, however, provides the means for a successful disambiguation. For if a causal inference drawn from (1) is not only based on the standard assumptions of BN-algorithms, as the causal Markov and faithfulness assumptions, but moreover on the assumption that e.g.

(P_I) I is an intervention variable for X with respect to Y

all ambiguities disappear. As we have seen above, (P_I), among other things, implies that I causes X. The only structure among the ones depicted in figure 2 in

[13] An expression as $Y \perp\!\!\!\perp Z \mid X$ is an abbreviation of $P(Y \mid Z \wedge X) = P(Y \mid X)$, which states that X screens off Y and Z.

which I in fact is a cause of X is structure (a). In (b), (c), and (d) I is an effect of X. That means if it is assumed that I has the interventionist properties laid out in (IV), the probabilistic data in (1) cannot stem from either (b), (c), or (d). It can, thus, unambiguously be determined to be the result of a causal structure such that I is a direct cause of X which, in turn, is a direct cause of Z and Y, i.e. of structure (a). All that is needed for this remarkable disambiguation is a seemingly unproblematic additional premise as (P_I).

The validity of this interventionist disambiguation is beyond doubt, as it essentially is of the following logical form:

> The empirical data recorded in (1) stems from causal structure (a) or (b) or (c) or (d).
>
> The data in (1) neither stems from (b) nor from (c) nor from (d).
>
> ---
>
> $\therefore$ The data in (1) stems from (a).

Adding (P_I) yields just one valid interventionist disambiguation of the causal analysis of (1). Supplementing either of the following interventionist premises disambiguates the causal modeling of (1) in two additional ways:

(P_Y) Y is an intervention variable for X with respect to I;
(P_Z) Z is an intervention variable for X with respect to Y.

The only structure in figure 2 that is compatible with both (1) and (P_Y) is (c), because only in (c) is Y a cause of X which is implied by (P_Y). For an analogous reason, the only structure that is compatible with (1) and (P_Z) is (d).

While all of these disambiguating arguments are clearly valid, their soundness crucially hinges on the truth of their second premises, which, as shown above, are consequences of the interventionist assumptions (P_I), (P_Y), and (P_Z), respectively. Accordingly, if (P_I), (P_Y), and (P_Z) are true, the pertaining interventionist disambiguations of the causal modeling of (1) are not only valid, but moreover sound. However, when applied to (1) these three assumptions are mutually exclusive. Or differently, no causal structure that could possibly generate (1) can satisfy more than one of (P_I), (P_Y), and (P_Z). That is, a new ambiguity emerges: If any, which of (P_I), (P_Y), and (P_Z) is true for the causal structure behind the four variables I, X, Y, and Z? Which of the three disambiguating arguments drawing on the interventionist framework is not only valid but moreover sound? Assessing the soundness of interventionist disambiguations of causal inferences inevitably calls for justifying the additional interventionist assumptions resorted to. Yet, if the pertaining interventionist methodology is embedded in a theoretical framework grounded in (M) and (IV), such justifications, as we have seen above, trigger infinite regresses, *viz.* justification regresses. Thus, just as strategy (i), (ii) merely replaces one type of regress by another.

At this point, the proponent of (ii) could respond that a narrowly conceived Euclidean project of causal reasoning only aims at reconstructing and assessing

the validity of causal inferences drawn on the basis of empirical data and supplementary causal assumptions without addressing the soundness of these inferences. If somebody is embarking on a project that exclusively aims to spell out what causal inferences are possible relative to what causal assumptions, he or she might be perfectly satisfied with simply reconstructing the three different interventionist disambiguations of the causal analysis of our exemplary data (1) as done above. The question of justifying the three mutually exclusive interventionist premises does not arise within such a context and, accordingly, no justification regresses are triggered either. However, such a narrow Euclidean project is rather limited in scope. It only yields what might be called a *logic of causal reasoning*. We have seen above that authors with sympathies for i-theories of causation, as Woodward (2003), Shapiro and Sober (2007) or Campbell (2007), do not have such a project in mind. They want to actually identify certain variables as causes of others on an i-theoretical basis. Whoever does not merely want to settle for valid causal inferences, but moreover is interested in the soundness of such inferences, i.e. interested in what causal structures *de facto* generated given data, cannot stop the i-theoretical regresses by merely assuming certain variables to have the interventionist properties laid out in (IV) and (M). Such assumptions call for stringent justifications which, in turn, cannot be provided within a conceptual framework that interdefines causation and intervention.

V Conclusion

We have seen that, even though an i-theory along the lines of (M) and (IV) can be argued to be conceptually informative, applying its interdefined conceptual fundament to concrete causal structures triggers infinite regresses of epistemic nature. It is impossible to ever conclusively identify a specific variable as an intervention variable within a theoretical framework grounded in (M) and (IV). These identification regresses can only be stopped if either prior causal knowledge is resorted to, or causal inferences based on (M) and (IV) are supplemented by assumptions to the effect that certain variables exhibit the interventionist properties laid out in (M) and (IV). Resorting to prior knowledge stops the identification regresses only at the expense of initiating other regresses: justification regresses. Simply assuming certain variables to comply with (M) and (IV) either only allows for a narrowly conceived Euclidean program of causal reasoning that exclusively assesses the validity of causal inferences, or also triggers infinite justification regresses. In sum, methodologies of causal discovery that are designed to output sound causal inferences in a finite number of steps cannot be based on an i-theory that interdefines causation and intervention as done by (M) and (IV). Or differently, there cannot exist an effective interventionist methodology that uncovers causal dependencies as defined by i-theories. Contrary to Woodward's own assessment and contrary to studies such as Campbell (2007) or Shapiro and Sober (2007), i-theories are not effectively applicable in contexts of causal discovery. At best, they provide

an analysis of the conceptual interdependence of causation and intervention. They do not, however, permit to ever actually identify a specific variable as a cause of another variable. I-theoretical interventionism may be argued to be conceptually unproblematic, but it may not be argued to be epistemically unproblematic as well. Suitably intervening on causal processes undoubtedly is a very powerful and efficient way of uncovering pertaining structures. For the very reason, however, that a highly causally loaded notion like that of intervention is of such central importance to uncovering causal structures, a methodology of causal reasoning cannot be based on an interventionist notion of causation as (M). It must either be based on a notion of causation that is introduced as a conceptual primitive, as done in p-theories, or draw on a conceptual analysis of causation that is independent of the notion of intervention.*

References

BAUMGARTNER, M. 2008a, 'Regularity Theories Reassessed', *Philosophia* **36**, pp. 327–354.

BAUMGARTNER, M. 2008b, 'Uncovering Deterministic Causal Structures: A Boolean Approach', *Synthese*. **URL:** *http://www.springerlink.com/content/x0487831qk67h455/*

BAUMGARTNER, M. forthcoming, 'Interventionist Causal Exclusion and Non-Reductive Physicalism', *International Studies in the Philosophy of Science*.

CAMPBELL, J. 2007, 'An Interventionist Approach to Causation in Psychology', in: A. Gopnik and L. Schulz, eds, *Causal Learning. Psychology, Philosophy, and Computation*, Oxford: Oxford University Press, pp. 58–66.

COLLINGWOOD, R. G. 1940, *An Essay on Metaphysics*, Oxford: Clarendon Press.

DE REGT, H. 2004, 'Review of James Woodward, *Making Things Happen*', *Notre Dame Philosophical Reviews*. **URL:** *http://ndpr.nd.edu/review.cfm?id=1455*

DOWE, P. 2000, *Physical Causation*, Cambridge: Cambridge University Press.

EBERHARDT, F., GLYMOUR, C. and SCHEINES, R. 2006, '$n-1$ Experiments Suffice to Determine the Causal Relations Among n Variables', in: D. Holmes and L. Jain, eds, *Innovations in Machine Learning*, Berlin: Springer, pp. 97–112.

GASKING, D. 1955, 'Causation and Recipes', *Mind* **64**, pp. 479–487.

GLYMOUR, C. 2004, 'Review of James Woodward, *Making Things Happen*', *British Journal for the Philosophy of Science* **55**, pp. 779–790.

HAUSMAN, D. 1986, 'Causation and Experimentation', *American Philosophical Quarterly* **23**, pp. 143–154.

HAUSMAN, D. 1998, *Causal Asymmetries*, Cambridge: Cambridge University Press.

HAUSMAN, D. and WOODWARD, J. 1999, 'Independence, Invariance and the Causal Markov Condition', *British Journal for the Philosophy of Science* **50**, pp. 521–583.

KVART, I. 1997, 'Cause and Some Positive Causal Impact', in: J. Tomberlin, ed., *Philosophical Perspectives 11: Mind, Causation, and World*, Atascadero: Ridgeview, pp. 401–432.

*I am grateful to Frederick Eberhardt, Jim Woodward, Delphine Chapuis, Richard Dawid, Isabelle Drouet, Mehmet Elgin, and John Norton for very helpful comments and discussions. Moreover, I thank the anonymous referees of *dialectica* for valuable comments on an earlier draft. Finally, I am indebted to the Center for Philosophy of Science of the University of Pittsburgh and to the Swiss National Science Foundation for generous support of this work (grant PBBE1 – 117031).

KVART, I. 2001, 'Causal Relevance', in: B. Brown, ed., *New Studies in Exact Philosophy: Logic, Mathematics and Science*, Oxford: Hermes Scientific Publications, pp. 59–90.

KVART, I. 2004, 'Probabilistic and Counterfactual Analyses', in: J. Collins, N. Hall and L. A. Paul, eds, *Causation and Counterfactuals*, Cambridge: MIT Press, pp. 359–386.

MACKIE, J. L. 1976, 'Review of G. H. v. Wright, *Causality and Determinism*', *Journal of Philosophy* **73**, pp. 213–218.

MENZIES, P. and PRICE, H. 1993, 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science* **44**, pp. 187–203.

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

SALMON, W. C. 1998, *Causality and Explanation*, Oxford: Oxford University Press.

SHAPIRO, L. and SOBER, E. 2007, 'Epiphenomenalism. The Dos and Don'ts', in: G. Wolters and P. Machamer, eds, *Thinking about Causes: From Greek Philosophy to Modern Physics*, Pittsburgh: University of Pittsburgh Press, pp. 235–264.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. 2000, *Causation, Prediction, and Search*, 2 edn, Cambridge: MIT Press.

STREVENS, M. 2007, 'Review of Woodward, *Making Things Happen*', *Philosophy and Phenomenological Research* **LXXIV**, pp. 233–249.

STREVENS, M. 2008, 'Comments on Woodward, *Making Things Happen*', *Philosophy and Phenomenological Research* **LXXVII**, pp. 171–192.

SUPPES, P. 1970, *A Probabilistic Theory of Causality*, Amsterdam: North Holland.

VON WRIGHT, G. H. 1971, *Explanation and Understanding*, Ithaca: Cornell University Press.

WOODWARD, J. 1997, 'Explanation, Invariance and Intervention', *Philosophy of Science* **64**, pp. S26–S41.

WOODWARD, J. 2003, *Making Things Happen*, Oxford: Oxford University Press.

WOODWARD, J. 2007a, 'Causation With a Human Face', in: H. Price and R. Corry, eds, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford: Oxford University Press, pp. 66–105.

WOODWARD, J. 2007b, 'Interventionist Theories of Causation in Psychological Perspective', in: A. Gopnik and L. Schulz, eds, *Causal Learning. Psychology, Philosophy, and Computation*, Oxford: Oxford University Press, pp. 19–36.

WOODWARD, J. 2008, 'Response to Strevens', *Philosophy and Phenomenological Research* **LXXVII**, pp. 193–212.