Against Interventions as Regulative Ideal: The Fragility of Nonlinear Feedback Relations

Abstract

It has long been recognized that nonlinear feedback systems are problematic for the interventionist account of causation. However, the discussion has stalled for two main reasons: insufficient clarity regarding the aims and scope of interventionism, and an insufficiently detailed analysis of nonlinear dynamics. This paper is a critique of the interventionist 'semantic' project, where ideal interventions elucidate what it means for one variable X to cause another variable Y. The behavior of the Lorenz system is discussed, and it is argued that ideal interventions do not and cannot reveal anything meaningful about the causal nature of nonlinear feedback relations.

KEYWORDS: Interventionism, Complexity, Idealization, Nonlinear Dynamics

INTRODUCTION

A decade and a half after its canonical formulation (Woodward 2003), the interventionist account of causation remains one of the most influential understandings of causation and causal explanation in the philosophy of science. Drawing on work in causal graph theory (Pearl 2009), it gives clear conditions how to analyze the meaning and structure of causal statements ('smoking causes lung cancer'). Cause and effect are represented by variables, and, roughly, the account tells us that the meaning of a causal relation is to be explicated in terms of 'ideal interventions', which, in their original 'surgical' variation (Woodward has recently allowed for non-surgical interventions: see later), first isolate the relation from the complex network in which it may be embedded, and subsequently reveal how change in the one variable occurs in virtue of change in the other variable.

Ever since its first introduction, interventionism has been dogged by objections that surgical interventions are not possible for many systems, especially complex systems with non-modular causal structures (Cartwright 2001, 2002, 2004, 2007; Mitchell 2008, 2009). However, such objections have never been perceived as fatal to the project. Interventionists have responded to such objections by arguing they based on a misunderstanding of what interventionism is trying to achieve: to account for what it *means* for one variable to cause another (Woodward 2003, Pearl 2009, Kuorikoski 2012). Interventionism is a semantic project, and it is perfectly permissible that the meaning of a causal relation be elucidated in ideal and even non-actualizable circumstances. Interventions, so it is claimed, are to be understood as a 'regulative ideal': an ideal methodology for evaluating the truth of causal claims concerning variables of a system (Woodward 2003, 2015).

In this paper I focus on the building block of complex systems behavior: nonlinear

feedback relations (NFRs). These relations are ubiquitous in nature and engineering systems, and their presence is increasingly recognized in sociological and economical systems (e.g. Guégan 2009 or Guastello 2013). After laying out some groundwork, I will consider some attempts to shoehorn NFNs into the interventionist framework, but will argue that in each case the resulting interventionist analysis fails the goals set out in the semantic project. Generalizing from this, I will propose that some causal relationships are 'fragile' against idealized intervention, and that interventions are not a regulative ideal for investigating such causal relationships. There is a broad swathe of causal reality about which interventionism has nothing to say.

I. Interventions in Complex Systems

In the interventionist framework, possible causes and effects are formally represented as variables, and the relation between cause and effect as a structural equation such as $Y = f(X)$. The variable X can be said to 'cause' Y if and only if an ideal intervention I on X, setting the value of X to x, leads to a change in the value of Y to $Y = f(x)$. The important philosophical work in this definition lies in the concept of an ideal intervention, and for I to be an 'intervention variable', it needs to meet a number of conditions (see Woodward 2003: 98-99), two of which are important for our purposes:

> $(I2)$ 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 X.
>
> $(I5)$ I does not alter the relationship between Y and any of its causes Z that are not on any directed path (should such a path exist) from X to Y. (Woodward 2003: 98-99)

One obvious problem here is that such surgical interventions are not possible in systems that have a non-modular causal structure: where (roughly) it is not possible to intervene on one causal relationship without affecting the nature of the other causal relationships in the same system. Nancy Cartwright was an early critic of this aspect of interventionism, and one of her oft-cited counterexamples is a carburetor, which is the component of a combustion engine that mixes fuel and air (Cartwright 2004). A carburetor involves many distinct causal processes, such as processes controlling airflow, air pressure, flow of gas, etc.; however, each process is dependent on the geometry of the chamber of the carburetor. This means that if any one of these causal processes were to be changed, the entire carburetor would need to be redesigned, thus changing the other causal processes. It is thus impossible to surgically intervene on any single part of the carburetor while both keeping the other parts fixed and preserving the essential function of carburetion (blending air and fuel). The point can be generalized to all systems where the components coordinate to perform a *function* and where the function simultaneously constrains the design of the components.

Another counterexample described by Sandra Mitchell (2008) is how some system maintain a certain output or behavior despite perturbations ('robustness'). For instance, genotype-phenotype maps often exhibit such robustness: intervening on some gene X may yet produce the *same* value of the trait Y, not because X has no causal impact on Y, but because the causal relation between some other gene Z and Y is retooled as to keep Y intact.

II. The Standard Response

The standard interventionist response to the problematic cases identified by Mitchell and Cartwright has consisted in pointing to the aims and scope of the interventionist project. The interventionist project aims at understanding the semantics or meaning of causal relations; however, whether an intervention can in fact be carried out is an "empirical question" that is the proper domain of scientific inquiry (Woodward 2010: n11). Interventionism is concerned with the "interpretative question" concerning causal relations, or in other words, with "what it *means* (in one perfectly good sense of "means") for X to cause Y or what one is *committed to* in claiming X causes Y." (Woodward 2015: 3587, original emphasis).

Applying this to Cartwright's and Mitchell's objections, the response is that, while adding layers of complexity to a system may make interventions physically impossible, insofar the additional complexity does not change the nature of the existing relations between the variables, these relations still can be elucidated by means of ideal intervention. Pearl makes a similar distinction between 'symbolic' (an equivalent term for 'ideal') and 'physical' intervention when responding to Cartwright's challenges:

> Surgery, and the whole semantics and calculus built around it, does not assume that in the physical world we have the technology to incisively modify the mechanism behind each structural equation while leaving all others unaltered. (...) Surgery is a symbolic operation. (Pearl 2009: 364).

A further nuance to this response is added by Kuorikoski (2012), who argues that, even if one grants Cartwright that it is not possible to intervene on the parameters of one subsystem of the carburetor without affecting the parameters of other subsystems, the

causal relationships within each subsystem – for instance between pressure and airflow – are still analyzable by means of ideal interventions. While it may not be possible to intervene on the parameters parameters of the structural equations, this is not a problem for the interventionist framework, which concerns only causal relations between variables (see also Woodward 2003: 327ff).

While an intervention need not be physically possible, this does not mean that any conceivable intervention is a candidate for elucidating the semantics of causal relationships. The condition placed on an intervention is that it be logically, conceptually and metaphysically possible to carry out (Woodward 2003: 112, 130-132). For instance, while changes in the sizes of the one non-perpendicular angles of a perpendicular triangle affects the size of the other non-perpendicular angle, such changes cannot be thought of as ideal interventions since the two angle sizes determine each other by virtue of mathematical necessity. By contrast, the connection between the gravitational constant and the Newtonian gravitational force is one of physical but not mathematical necessity, and hence can be subject to ideal intervention. In sum, surgery may be a symbolic operation, but it is not allowed to operate on the symbolic relations that represent logical, conceptual, or metaphysical necessities.[1]

The key question considered in this paper is whether the standard response actually delivers. Does it succeed in shielding the semantic project of interventionism by relegating the worries of Cartwright and Mitchell to the realm of 'empirical questions'? I will argue it does not, and that complex systems (and, more specifically, the causal relations that constitute them) form a more fundamental problem for the interventionist project

[1]This move could be subjected to critique from many sides. For instance, is it really so that physical and mathematical necessity can be neatly distinguished from each other? Or, if interventions cannot be carried out on non-causal relations (such as mathematical or metaphysical relations, does this not imply a circularity? I will leave these worries aside in this paper (cf. Reutlinger 2012).

than realized.

III. Regulative Ideals and Semantics

In this section, in order to lay the ground for the critique of interventionism, we need to analyze in more detail what precisely regulative ideals are, and what precisely the semantic project of interventionism is. To this end it will be useful to encapsulate the core concepts by means of some minimal formalization. According to interventionism, the meaning of 'X causes Y' is given by $\mathscr{I}$, where $\mathscr{I}$ is the set of intervention counterfactuals:

$$\mathscr{I}_{X\to Y} = \{do(X = x) \to Y = y| \text{ for all } x\}$$

where the operator $do(\cdot)$ represents an ideal intervention on the variable X, and where X is allowed to vary along some relevant range of values (see Pearl 2009). Thus the counterfactuals in $\mathscr{I}$ do not necessarily correspond to the counterfactuals obtained by passive observation of the behavior $\mathscr{B}$ of $X \to Y$:

$$\mathscr{B}_{X\to Y} = \{X = x \to Y = y| \text{ for all } x\}$$

These counterfactuals include effects from other variables Z, and so do not necessarily indicate the true quantitative nature of the causal relation between X and Y.

The set of intervention counterfactuals $\mathscr{I}$ is obtained by a series of ideal experiments (or ideal interventions) on the causal relation $X \to Y$, and the core interventionist claim is that this set $\mathscr{I}$ specifies exhaustively what we mean when we say 'X causes Y'. The true quantitative nature of the causal relation between X and Y is given by $\mathscr{I}$, not $\mathscr{B}$.

For this reason, ideal interventions can be thought of as a regulative ideal that show

how, through ideal experimentation, the semantics of a causal relation is to be established – and thus how $\mathscr{I}$ is to be distinguished from $\mathscr{B}$:

> The notion of an intervention (...) represents a regulative ideal. Its function is to characterize the notion of an ideal experimental manipulation and in this way to give us a purchase on what we mean or are trying to establish when we claim that X causes Y. (Woodward 2003:130)

In the following two subsections I will distinguish between two separate but related semantic functions of a regulative ideal, as well as two types of regulative ideal.

1. Causal Representation and Causal Discovery

There are two ways of interpreting the set of intervention counterfactuals $\mathscr{I}$, each indicating a different semantic project. The first way is to interpret $\mathscr{I}$ as the ideal representation of a causal relationship. In particular, $\mathscr{I}$ defines the function $f : X \to Y : x \to y$ that in turn allows for the definition of a structural equation:

$$Y = f(X)$$

The fact that the relation between two variables can be represented as a functional relationship is non-trivial and philosophically significant, because it implies that a well-defined value of the cause variable can be associated with a well-defined value of the effect variable (or with a well-defined probability that the effect variable would take some value). The probabilistic version of the structural equation above is $P(Y) = f(X)$ for some probability function P. In this case, while the same value of X could lead to different values of Y, the probability of Y taking a certain value is fixed given X.

This representation of the causal relation in a structural equation in turn allows for representation by means of a *directed causal graph*, where where an arrow indicates a direct causal relationship. These causal relationships allow for relations of conditional independence to be read off of the graph according to the criterion of 'd-separation' (due to Geiger et al. 1990). In intuitive terms, X 'd-separates' Y and Z if and only if any path that connects Y to Z, crosses X. This is a graphical representation of a particular independence relation: if X is held fixed, then Y and Z are probabilistically independent, or equivalently, the correlation between Y and Z conditionalized on $X$is zero.

Finding the right causal graph with the right properties (such as d-separation) is an important part of the work in the Bayesian nets literature, and also has received a lot of attention in the case of feedback equations, since it is a nontrivial matter to find a causal graph that adequately represents the causal nature of some feedback relations (see Clarke et al. 2013, Gebharter and Schurz 2016). The purpose of this work can be understood as subsuming nonlinear feedback relations under ideal causal representation.

By contrast, this paper is focused on the second interpretation of $\mathscr{I}$: as representing an ideal method of causal discovery. It represents the body of information that a scientist would discover if disencumbered of the constraints of physical possibility. For instance, if the physicist set out to test the relation between the gravitational constant g and the strength of the gravitational force F_g, she would ideally manipulate g with respect to F_g and would discover a set of counterfactuals $\mathscr{I}_g$ that showed a linear proportionality between the two variables. The set $\mathscr{I}_g$ would be the result of her ideal causal discovery, and could be used to specify what it meant for g to cause F_g.

Note that an 'ideal' method of discovery does not imply the most preferable or most efficient method of discovery. It may not even be a possible method of discovery – as is manifestedly the case with $g \rightarrow F_g$. Calling it 'ideal' refers to how interventions

regulate how we intend the semantics of causation. In other words, to say 'g causes F_g' means that, if it were possible, we would intervene on g with respect to F_g and observe the resulting counterfactuals. This is also how we would ideally test the proposition 'g causes F_g'.

Ideal discovery and ideal representation are two sides of the same coin, and can be thought of as two epistemic functions that a regulative ideal can play. A regulative ideal can express an ideal methodology – an ideal process for getting at the meaning of a causal relation – or it can express the ideal outcome of that methodology, where the relations between different variables can be visually represented in a directed graph and quantified by a set of structural equations.

2. Constitutive and Merely Regulative Ideals

It may seem trivial to say that causal discovery should yield some predictive value as to real-life behavior. The set of intervention counterfactuals $\mathscr{I}_{g \to F_g}$ should contain some predictive value, perhaps in combination with other sets of idealized interventions on other variables, as to how g and F_g related in real life, outside of the idealized experimental setting. After all, if there was zero predictive value, how could we know the counterfactuals resulting from ideal interventions would not be fictions?

This doubleness of a regulative ideal can be illustrated with the metaphor of a map, which both represents the landscape and guides the traveler through it. If a map does not guide a traveler accurately through the landscape then the map is equivalent to a drawing, a fiction – or worse, since misplaced confidence in a fake map can have disastrous consequences. A map is not a pure fiction when it contains some information about the lay of the land, i.e., when the signs on the map can be used for purposes of predicting

the features of actual geography.

Likewise, if a regulative ideal such as an ideal intervention is not to be some artificial construct (or *merely regulative*), it needs to yield some predictive information of the causal behavior – in other words, it needs to be *constitutive* of how the causal relation actually behaves in reality.

We will make this distinction between constitutive and merely regulative ideal more accurate still, but consider first how the distinction cuts to the heart of the interventionist project. Here is a remarkably similar claim endorsed by Pearl:

> The effect of implementing such a complex policy can be predicted using the 'surgical' semantics of the do-calculus in much the same way that properties of complex molecules can be predicted from atomic physics. (Pearl 2009: 363).

The results of the surgical interventions (i.e., $\mathscr{I}$) can be used as building blocks to reconstruct the actual behavior of causal relations and even whole causal networks.

Interventions (whether surgical or not: see later) generate a set of intervention counterfactuals $\mathscr{I}_{X\rightarrow Y}$ that is an *idealization* of the relationship between X and Y as found it in reality. Ideal interventions perturb the causal network in which X and Y find themselves, and the interventionist assumption is that such intervention does not significantly distort the causal relation between X and Y, so that $\mathscr{I}(X \rightarrow Y)$ can be used to explain how changes in X actually cause changes in Y. This allows $\mathscr{I}$ to be explanatory of the actual causal relation, and not some merely artificial representation.

A minimal sense which *mathcalI* can be understood to be explanatory is to be predictive to a certain degree of $\mathscr{B}$. This however does not mean that a single set *mathcalI* should unambiguously determine $\mathscr{B}$. For instance, a single set of interventions may

yield $\mathscr{I}_1$ which could correspond to multiple causal models (part of the same 'interventional Markov equivalence class': see Eberhardt and Scheines 2007) and thus multiple causal behaviors. Additional sets of interventions[2] on other variables, yielding further sets of intervention counterfactuals $\mathscr{I}_2$, $\mathscr{I}_3$, and so on, may be necessary to narrow down the range of possible causal models, and thus to reconstruct the actual causal behavior.

One way to precisely think about how intervention counterfactuals can explain actual behavior – and thus the contrast between constitutive and merely regulative ideal – is by means of mutual information I (not to be confused with an intervention variable):

$$I(\mathscr{I};\mathscr{B}) = H(\mathscr{B}) - H(\mathscr{B}|\mathscr{I})$$

$I(\mathscr{I};\mathscr{B})$ is high when knowing $\mathscr{I}$ strongly reduces the uncertainty over possible values of $\mathscr{B}$. It is important to note that this measure only gives a precise operationalization when the probability distributions over the possible behaviors $\mathscr{B}$ or possible sets of intervention counterfactuals $\mathscr{I}$ is known exactly. This is of course unrealistic in most situations; nonetheless, the measure of mutual information is useful in order to explicate explanatoriness as reducing the range of possible outcomes (in this case, possible behaviors $\mathscr{B}$). In this sense mutual information can be understood as a measure of predictiveness.

In the following I will consider nonlinear feedback relations, and will argue that no matter how these are analyzed by means of ideal interventions, that a set of intervention counterfactuals $\mathscr{I}$ is obtained that contains little to no predictive information about the actual causal behavior.

[2]Under certain conditions, $N-1$ sets of interventions may be required for a system with N variables: Eberhardt, Glymour, and Scheines 2006.

IV. NONLINEAR FEEDBACK RELATIONS

Nonlinear feedback relations (NFRs) are by no means rare occurrences in nature, and are present in systems ranging from chemical reactions and metabolism to weather and financial markets. Feedback relations are often embedded within a larger network, termed here nonlinear feedback networks (NFN), of which a minimal example is given by Figure 1. Such a schema may represent gene-phenotype mapping, where X and Z may represent two separate genes, both of which affect the expression of phenotypic trait Y, and both of which affect the expression of the other. Or it may represent a market for some commodity, where X represents the demand for a commodity, Z represents the price, and Y may represent the supply. At market equilibrium, price and demand are related by negative feedback, but positive feedback can also occur, as during market bubbles.

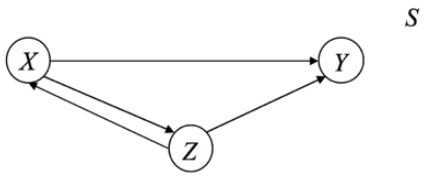


Figure 1: Minimal nonlinear feedback network.

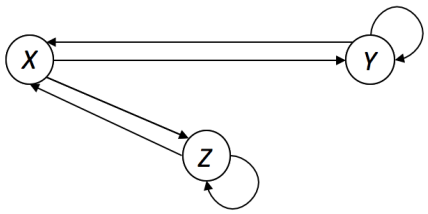
NFRs are typically left undiscussed in the philosophical literature on interventionist causal discovery (cf. Eberhardt and Scheines 2007; Eberhardt, Glymour, and Scheines 2006). In the wider scientific literature, causal discovery in general model space is an active area research, but even there the analysis beyond linear acyclic networks is limited to either nonlinear relationships (Hoyer et al. 2009) or linear cyclic networks (Richardson and Spirtes 1999; Hyttinen et al. 2013; Triantafillou and Tsamardinos 2015; Lacerda et al. 2008).

In the literature on causal representation, nonlinear feedback has received more attention, but it has proved to be a difficult test case. The efforts to represent linear feedback cycles in causal graphs with d-separation have been successful (Spirtes 1995, Koster 1996, Richardson and Spirtes 1999); however, generalizations to nonlinear feedback cycles, even with further limiting assumptions, have been decidedly more mixed (see for instance the result in Pearl and Dechter 1996, later shown to be false by Neal 2000).

In the recent philosophical literature on causal representation, nonlinear feedback has received attention regarding how it can even be represented as a causal graph at all. Some have been optimistic about the endeavor and have advanced formalisms (Clarke et al. 2004; Gebharter and Schurz 2016), whereas others have been more skeptical (Weber 2016, Kaiser 2016). This philosophical literature will inform the discussion in the next section, but to orient the reader, it is important to emphasize again that the present paper primarily concerns causal discovery, and thus has slightly different aims from the contributions in this literature.

As a test-case for the rest of the paper, we will consider how the Lorenz system of equations describes how changes in atmospheric pressure X affect the vertical temperature gradient Y:

$$\text{(LS)} \qquad \begin{aligned} \tfrac{dX}{dt} &= \sigma(Z-X) \\ \tfrac{dY}{dt} &= XZ - bY \\ \tfrac{dZ}{dt} &= X(r-Y) - Z \end{aligned}$$



where Z represents the horizontal temperature gradient. The parameters σ, r and b are

real and positive, and describe various physical properties of the fluid.[3]

It is clear that X, the rate of convection, causally affects Y, the vertical temperature gradient, and vice versa. However, X correlates with Y is affected by a third variable Z, representing the horizontal temperature gradient. how The question we set ourselves is whether this causal relationship can be accurately discovered and described through ideal interventions. To that end, for the remainer of this section I will outline how the behavior of $X \rightarrow Y$ is described in nonlinear dynamics (simplified from Sparrow 1982): in terms of topological properties such as equilibria, periodic orbits, and strange attractors.

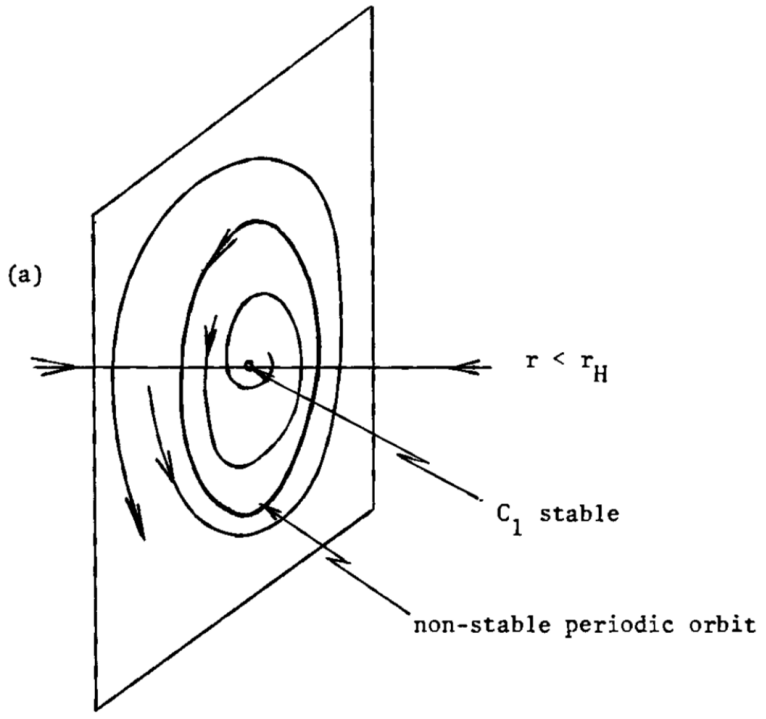


Figure 2: Convergence to Stable Equilibrium; Periodic Orbit; Divergence from Orbit (Reproduced from Sparrow 1982: 12).

[3] σ is proportional to the Prandtl number – defined as the ratio of viscous diffusion rate and thermal diffusion rate – r is proportional to the Rayleigh number – which can roughly be understood as a measure for how much heat in the liquid is transferred by convection – and b is a constant describing the geometry of the space occupied by the liquid.

(1) *Stable equilibrium* X and Y are in stable equilibrium when changes in X have only a short-lived effect on Y, and Y quickly relaxes back to its previous value. The stable value of Y is called the 'equilibrium' value. A perfect stable equilibrium can be represented by $\mathscr{B}(X = x \to Y = y_{eq})$ where all values of X are projected onto a single equilibrium value $Y = y$. In the Lorenz system, the origin $(X = 0, Y = 0, Z = 0)$ is a stable equilibrium for $r < 1$, and is globally attracting, meaning that, regardless of the initial value of (X, Y), the asymptotic value is $(0, 0)$.

(2) *Periodic orbits* X and Y are in a periodic orbit when when $\mathscr{B}(X = x(t) \to Y = f(x(t)))$, for a parametrization variable t and a smooth function f which maps onto itself with a certain period. Thus changes in X generate an well-defined oscillation in in Y.

In the Lorenz system, periodic orbits can occur when $r > 1$ (see Figure 2), and they orbit one of two equilibrium points that appear when $r > 1$ (see also discussion section). These orbits can be themselves considered stable or unstable (the periodic orbit in Figure 2 is stable in one direction, and unstable in the other).

(3) *Strange Attractors* When are X and Y have been captured by a strange attractor, changes in X cause changes in Y, and while the pair (X, Y) does not converge on a stable equilibrium or a periodic orbit, but rather converges to (or remains confined to) a limited region. The region contains an uncountable infinity of aperiodic orbits (Sparrow 1982: 21), such that two aperiodic orbits arbitrarily close may be arbitrarily far apart after some time t. The limited region is called a 'strange attractor' for this chaotic behavior.

More precise characterizations of strange attractors either make use of the topological notion of fractal[4], or the dynamical notion of sensitivity to initial conditions (and Lyapunov exponents). While the former is more fundamental, the latter is more useful

[4]The intersections of the paths within a strange attractor with the xy-plane resemble Cantor dust, and an infinite path followed within the attractor has a dimension larger than that of a plane, and smaller than that of a volume (≈ 2.06, see Viswanath 2004).

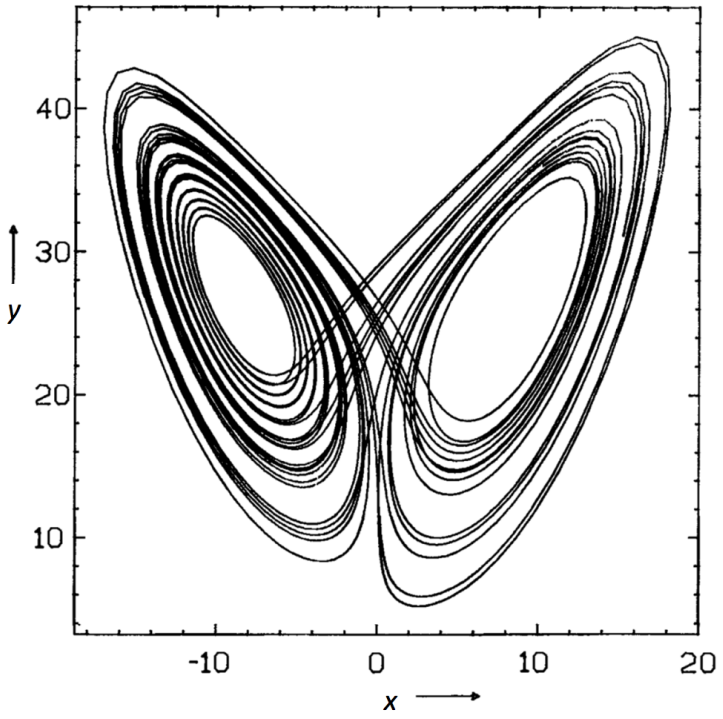


Figure 3: Strange Attractor (Reproduced from Sparrow 1982: 2).

in present context.

Note that this list is non-exhaustive, and that composite behaviors occur, such as temporary capture by an attractor-like invariant set followed by convergence to stable equilibrium (also referred to as 'preturbulence', see Figure 4), or the combination of multiple quasi-periodic orbits (see Sparrow 1982).

V. Interventionist Causal Discovery for NFNs

There is a vast array of methods of causal discovery in the statistical inference and signal processing literature (see e.g. Kay 1993). For our purposes we need only consider how the ideal causal discovery of feedback relations has been approached in the interventionist framework. In the following we will be starting from causal relations which

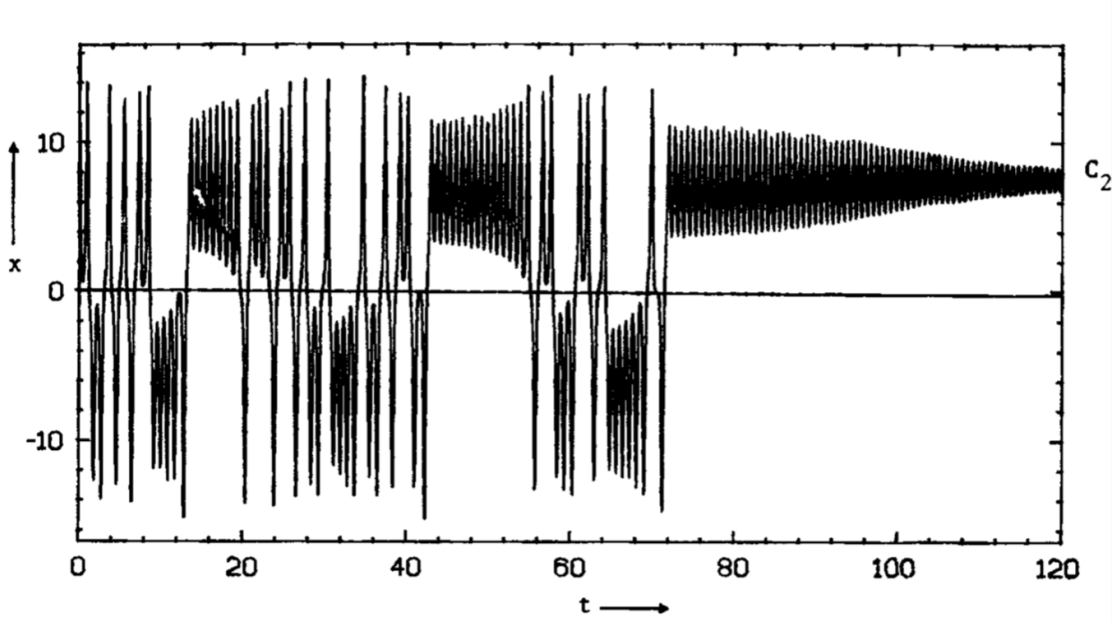


Figure 4: Preturbulence (Reproduced from Sparrow 1982: 31).

are then 'prepared' (idealized) so that ideal interventions can be carried out.[5] The key question then is whether the resulting sets of intervention counterfactuals are predictive of the (known) real causal behavior, or whether they are more properly attributable to the modified or idealized causal relation.

1. Analysis by Brute Simplification

The analysis originally advocated by Woodward (Woodward 1999) – called here the analysis by 'brute simplification' – involved first idealizing the feedback causal relation as a unidirectional relation, and then carrying out the ideal intervention $I_X \rightarrow X \rightarrow Y$. The system Woodward considered (Woodward 1999: 234) was the following linear feedback system (see also Spirtes 1995 and Koster 1996):

[5]The problem of how to prepare a system for ideal intervention where the causal relations are unknown will be left aside.

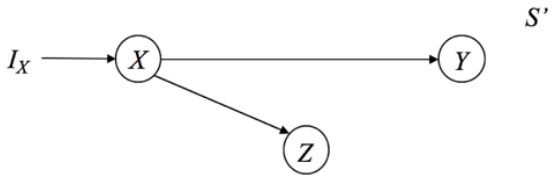


Figure 5: The idealized Lorenz system, receptive for ideal interventions on X with respect to Y.

$$X_3 = aX_1 + bX_4 + U$$

$$X_4 = cX_2 + dX_3 + V$$

His proposed solution was to set one of the variables in the causal feedback to a fixed value $(X_3 = k)$, while keeping the causal dependence of X_4 on X_2 intact, thus allowing for a free manipulation of X_4 by modulating the value of X_2. This solution is echoed by Pearl in response to Cartwright (Pearl 2009: 364-365).

Would it work for a NFN as in Figure 5? An idealized manipulation I of X with respect to Y could not (as per $(I5)$ mentioned in section 1) affect Y by any other pathway than $X \rightarrow Y$ and would have to (as per $(I2)$) fix X by the intervention I alone. So in effect, the idealized manipulation would need to be carried on the network represented in Figure 2, where Z does not affect Y and does not affect X.

However, the question is whether the resulting set of intervention counterfactuals $\mathscr{I}$ contains any information about the behavior $\mathscr{B}$ of the variables. Consider as an example how this strategy applies to the Lorenz equations. First, we fix Z to a constant C_1 since wish to isolate the causal relation between X and Y (this is $(I5)$). Second, since we want the ideal intervention alone to affect X (this is $(I5)$), so this means that Y cannot affect X, and X cannot affect itself (so $dX/dt = 0$. This yields the following idealization of the

Lorenz equations:

$$\frac{dY}{dt} = X(r - C_1) - Y.$$

Such an equation is easily integrable (see Robinson 2004) and yields following set of intervention counterfactuals $\mathscr{I}_S = \{X = x \rightarrow Y = X(r - C_1) + C_2 e^{-t} | C_2 \in \mathbb{R}\}$, indicating a linear relationship between X and Y. Is this set of counterfactuals a good candidate for describing the causal nature of $X \rightarrow Y$? The problem is that knowing the result of the ideal experimentation, $\mathscr{I}_S$, does not help in narrowing down the range of possible behaviors associated with X and Y. No combination of interventions on other variables, similarly following the brute simplification approach, will be able to inform us about the actual behaviors of X and Y, which we know from other methods and which range from approach to stable equilibrium to chaos. In other words, the predictiveness of $\mathscr{I}_S$ ($I(\mathscr{I}_S; \mathscr{B}_{X \rightarrow Y})$) is very low. Based on this, one can argue that $\mathscr{I}_S$ does not represent the causal nature of $X \rightarrow Y$ at all, but rather the idealized preparation of $X \rightarrow Y$ obtained through brute simplification.

Gebharter and Schurz (2016) echo Woodward's brute simplification approach when conjecturing that 'effects of interventions for cyclic graphs possibly featuring bidirected arrows can be computed as usual'. They propose that, by first deleting all arrows pointing at X (doing this to the Lorenz system in effect yields Figure 5), one can use the $d-$separation information on that graph to compute some set of intervention counterfactuals $\mathscr{I}$[6] (Gebharter and Schurz 2016: 940-941). The problem with this conjecture is not that it's false: it is true that $\mathscr{I}$ (or $P(Y|\hat{X})$) can be computed in this fashion. The problem is that the resulting $\mathscr{I}$ is an abstraction that has no resemblance to the real

[6]Equivalent to $P(Y|\hat{X})$ in their notation, where $\hat{X}$ is the manipulated variable.

behavior $\mathscr{B}_{X\to Y}$ (or $P(Y|X)$) as evidenced by the unmanipulated Lorenz system.

In summary, the analysis by brute simplification of a NFR $X \to Y$ abstracts away from the feedback relation, yielding a set of intervention counterfactuals $\mathscr{I}$ that is of little to no predictive value concerning the behavior of the actual system, even in combination with similar interventions on other variables. The idealization required for carrying out interventionist analysis distorts both the system as well as the causal relation $X \to Y$, and what ends up being described is an abstraction – the 'prepared' causal relation – rather than the actual causal relation.[7]

2. Dynamic Bayesian Networks

Instead of simply replacing feedback relationships by unidirectional ones, another method is prepare the causal relationship $X \to Y$ and the embedded network by time-indexing the variables and replace the feedback relationship by a unidirectional one for only a short period of time. The main idea is that feedback between X and Y means that an intervention on X first induces a change in Y, which then induces a change on X, and so on. In this way a nonlinear feedback relationship can be represented by a directional acyclic graph that have a temporal dimension, termed dynamic Bayesian networks.

Some have proposed approaches where some bidirectional causation between two variables is retained in the time-indexed graph (when due to a common variable: see the 'dynamical causal model' approach in Gebharter and Schurz 2016). For purposes here[8] I will consider only how nonlinear feedback relations are analyzed as dynamic Bayesian

[7] Eberhardt and Scheines 2007 note the possibility of 'soft' interventions, where a variable is intervened on but its other causes are allowed to influence the variable, to establish the presence or absence of a causal connection. Their basic idea – intervention without surgery – is subsumed by the approach described in the next section.

[8] The bidirectional causal relationships in, for instance, the Lorenz equations are not due to common causes, and thus cannot be replaced by unidirectional causal relationships at a finer grain of analysis.

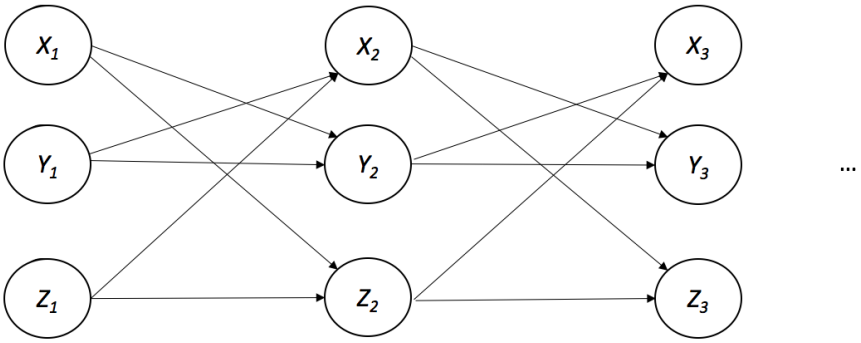


Figure 6: The Lorenz system unrolled over time.

networks (following Clarke et al. 2014), but the conclusions apply more generally to time-indexing approaches. Constructing a dynamic Bayesian network to represent, for instance, the Lorenz equations is relatively straightforward and would look something like Figure 6.

However, the question at stake here is not whether the Lorenz equations can be represented as a DBN (they eminently can, because this is how they are computed numerically in simulations), but rather whether DBNs help support the ambitious semantic claim of interventionism: every causal relation, insofar it is causal and not mathematical or conceptual, can be elucidated by means of intervention counterfactuals. Can the causal relation $X \rightarrow Y$ in the Lorenz system, insofar it is a causal relation, be elucidated by means of ideal interventions on the DBN?

To see how ideal interventions are carried out on the DBN, consider how first the differential equations can be replaced by difference equations (setting the time-step $\Delta t =$

1):

$$
\begin{aligned}
& & X_{t+1} &= X_t + \sigma(Y_t - X_t) \\
& (\mathrm{LS}*) & Y_{t+1} &= Y_t + (X_t(r - Z_t) - Y_t) \\
& & Z_{t+1} &= Z_t + (X_t Y_t - bZ_t)
\end{aligned}
$$

These difference equations (LS*) represent a countable series of structural equations, quantifying the DBN represented in the graph in Figure 6.

Once the network has been unrolled in this fashion, the question still remains how one is to intervene on it in order to discover the 'correct' intervention counterfactuals $\mathscr{I}_{X \to Y}$ (if there are any to be had: see discussion section). Here I will consider the possibility of relaxing requirement $(I5)$, allowing for non-surgical interventions on X where other causes of Y (i.e., Z) are allowed to vary (cf. Eberhardt and Scheines 2007, Woodward 2015). This creates a range of possible strategies of intervening, each with a different set of counterfactuals $\mathscr{I}_{U_1}$, $\mathscr{I}_{U_2}$, and so on.

The first possibility, where X_t is surgically intervened on with respect to Y_{t+1}, reduces to the the brute simplification intervention method previously discussed, where the resulting intervention counterfactuals $\mathscr{I}_{U_1}$ describe a linear relationship, with little to no predictive value concerning the actual causal behavior $\mathscr{B}_{X \to Y}$.

At the other extreme, a set of interventions could simply map each causal relationship $X_t \to Y_{t+1}$, $X_t \to Z_{t+1}$, and so on. Thus a vast sequence of sets of intervention counterfactuals would be obtained, mapping each causal arrow in Figure 6: $\mathscr{I}_{U_2} = \{\mathscr{I}_{X_t \to Y_{t+1}}, \mathscr{I}_{X_t \to Z_{t+1}}, \mathscr{I}_{Y_t \to X_{t+1}}, \mathscr{I}_{Y_t \to Y_{t+1}}, \mathscr{I}_{Z_t \to Z_{t+1}}, \mathscr{I}_{Z_t \to X_{t+1}}, ...\}$. The problem here is not that it delivers too little, but too much. The $\mathscr{I}_{U_2}$ above represents how each variable can *possibly* affect the other variable, at each moment in time, and so does not tell us specifically

how X affects Y, but rather is just a reproduction of the causal behavior of the entire causal system: $\mathscr{B}_{X\to Y}$, $\mathscr{B}_{X\to Z}$, $\mathscr{B}_{Z\to X}$, and so on.

In similar fashion, other ways of intervening on (LS*) either reproduce the causal behavior of $X \to Y$ or yield intervention counterfactuals that are of little to no predictive value. For instance, one could intervene on the initial value of X, setting X_0 to a fixed value, and subsequently observing the changes in Y. One would obtain $\mathscr{I}_{U_3} = \{do(X_0 = x) \to Y_1 = y_1, Y_2 = y_2, \dots\}$. This is in fact the way to determine the structure of paths followed by (X,Y) – for instance, what the limits of the basins of attraction are. However, here the problem is that $\mathscr{I}_{U_3}$ incorporates information about how Z and subsequent values of X affect the subsequent values of Y – not just how changes in X_0 affect Y. Hence the resulting counterfactuals do not describe the specific causal relation $X \to Y$.

Finally, one could carry out a fat-hand intervention on multiple variables, setting $(X_0 = x, Y_0 = y, Z_0 = z)$, and subsequently observing the changes in Y: $\mathscr{I}_{U_3} = \{do(X_0 = x, Y_0 = y, Z_0 = z) \to Y_1 = y_1, Y_2 = y_2, \dots\}$. However, again here $\mathscr{I}_{U_3}$ simply reproduces an aspect of $\mathscr{B}$, the causal behavior of the Lorenz system, and cannot be considered as the semantics of $X \to Y$.

Have I not put the bar artificially high for the interventionist approach, setting up the problem in a way it cannot possibly resolve? I will consider such objections in the final section, but it suffices for the moment to point out that the interventionist project has the ambition to analyze *all* causal relations: insofar $X \to Y$ is a causal relationship, and not a mathematical, conceptual, or metaphysical necessity, its meaning can be unambiguously given by ideal interventions. It should not matter whether $X \to Y$ is linear relationship and in relative isolation from other causal influences, or nonlinear and embedded within a larger network.

3. The Fragility of Causal Relations

The preceding discussion can be encapsulated by a more general argument that makes the distinction between causal relations that are *fragile* and 'break' under ideal intervention – and thus that are not amenable to interventionist analysis –, and those which are robust against ideal intervention.

Let X and Y be causally related in a nonlinear feedback relationship, and let X and Y be embedded in a wider network S with at least one other variable Z which is also related to X by means of a nonlinear feedback relationship. Then an ideal intervention (or a series of ideal interventions) on X with respect to Y will do one of two things: either it will keep all other causes of Y (i.e., Z) constant (a 'surgical' intervention) or it will not.

If the former, then S is idealized to an idealized network S' by replacing feedback relations $X \rightarrow Y$ and $X \rightarrow Z$ by unidirectional ones $X \rightarrow Y$ and $X \rightarrow Z$. Hence, in the resulting intervention counterfactuals $\mathscr{I}_A$, Y only depends on the value of X, and not on the value of Z. The actual causal behavior $\mathscr{B}$ of $X \rightarrow Y$ is highly sensitive to values of Z, and therefore $\mathscr{I}$ contains no predictive information concerning $\mathscr{B}$.

If soft interventions are carried out, then the interventions on X also allow for Z to vary and to influence Y. The resulting intervention counterfactuals $\mathscr{I}_B$ duplicate the causal behavior $\mathscr{B}_{X \rightarrow Y}$, or an aspect of it.

Therefore it is impossible to both isolate the causal impact of X on Y and for the resulting intervention counterfactuals to have any predictive value of how changes in X actually impact changes in Y.

The causal relation $X \rightarrow Y$ can be said to be *fragile* in face of ideal interventions: if an attempt is made to isolate $X \rightarrow Y$ – even under ideal circumstances when no physical

intervention is actually carried out – the causal relationship is 'broken' in the sense that we are left with something ($\mathscr{I}$) that is a fiction and does not resemble reality.

For such a fragile causal relationship, an ideal intervention is either too hard – yielding counterfactuals that are mere abstractions – or too soft – yielding counterfactuals that could be obtainable by passive observation. Ideal interventions cannot elucidate the semantics of such causal relationships – fragile relationships remain opaque to interventionist analysis.

VI. Discussion and Objections

1. Unanswerable causal questions

A possible objection is that for the causal relation $X \rightarrow Y$, there is simply no meaning to be had. There is no 'correct' set of intervention counterfactuals, and that is because we are not committed to anything in particular when, for instance in the case of the Lorenz equations, we say that changes in the rate of convection cause changes in the vertical temperature gradient. Therefore the case considered in this paper is not a problem for the interventionist approach.

One way in which this objection can be further precisified is terms of what Woodward terms 'unanswerable causal questions' (2015: 3592 ff). The example Woodward considers (drawn from Angrist and Pischke 2009) is whether starting school at age seven instead of six improve educational outcomes. This question would seem amenable to interventionist analysis (i.e., vary the starting school age, and then observe the educational outcomes). The complicating factor here is that children age seven automatically have greater maturity (e.g. more experience, more brain development) than those age six, and this fact alone may improve educational outcomes. So what would need to be done

would be to manipulate the starting age while keeping maturity constant: a mathematical impossibility. So, Woodward concludes, what seemed like a legitimate topic for interventionist inquiry turned out to depend on a mathematical identity and so is in fact unanswerable. Upon reflection, it turned out we are not committed to anything when we might claim that starting school at age seven instead of six improves educational outcomes.

This is version of the objection is easily answerable, since the test-case considered in this paper did not involve a mathematical relationship falsely portrayed as a causal relationship. However, a general strategy is suggested: can the interventionist claim, in some way, that the causal relation $X \rightarrow Y$ considered in this paper is ambiguous, and when disambiguated is no longer a problem? The next objection follows this general strategy.

2. System-level versus relation-level interventions

In this paper I have moved back and forth between 'changes in X cause changes in Y' and 'changes in X cause changes in Y *embedded* in the Lorenz system'. Was this legitimate? In particular, one could object that I set ideal interventions an impossible task: they were to elucidate the meaning of 'X causes Y' in isolation from the effect of other variables *as well as* yield a set of counterfactuals that was predictive of the actual behavior of 'X causes Y', which includes the effect of other variables. In reality, there are two questions here: the system-level and the relation-level causal behavior, and there are different types of interventions suited for each (surgical versus fat-hand), but no single type of intervention can elucidate both 'changes in X cause changes in Y' and 'changes in X cause changes in Y *embedded* in the Lorenz system'.

This objection can be defused by showing it rests on a confusion. Consider what are we committed to when we say 'in absence of friction, changes in the inclination of the plane causally affect changes in the rate of acceleration of the object on that plane'. There are certain counterfactuals that describe this causal relationship, and these counterfactuals are explanatory (in the sense of predictive) of the actual causal behavior of an object on a plane (with friction). There are certain key differences – for instance, the object will only start to move once the inclination reaches a certain critical angle (because of static friction) – but once the force of gravity overcomes static friction, then the counterfactuals describing the frictionless behavior are predictive, given some corrective factor, of the actual causal behavior.

The causal relation between 'inclination of the plane' and 'rate of acceleration' are robust against idealized interventions. Idealization distorts the causal structure of the system, but the resulting counterfactuals are explanatory (predictive) of the real behavior. However, this is not the case with the relation between convection and vertical temperature gradient, which is fragile against idealized interventions. An idealized intervention on convection with respect to the vertical temperature gradient is of little to no explanatory value for the actual causal behavior.

Thus the confusion underlying the objection lies in thinking that 'changes in X cause changes in Y' and 'changes in X cause changes in Y *embedded* in the Lorenz system' are two separate and independent questions. They are not: the first is an idealization of the second. The second causal statement cannot be analyzed by means of surgical interventions. By contrast, the first can be, but at the price that the resulting counterfactuals are of little to no explanatory value for the actual causal behavior. The ideal of intervention is shown to be a merely regulative ideal, not constitutive.

The lesson to be drawn from this failure of interventionism is that some causal rela-

tions cannot be elucidated by means of intervention counterfactuals. Can they be elucidated by other means? I do not think so. I do believe it is a relatively pointless question to ask how convection affects the vertical temperature gradient while keeping the horizontal gradient constant. There are more meaningful ways of investigating the causal relationship than the strict interventionist approach. However, the point remains that we are at the limits of interventionist approach to causation. Here we have two variables X and Y in a causal relationship, and with ideal interventions unable to say anything meaningful about the nature of that causal relationship. This area of causal reality is *opaque* to interventionist analysis.

3. Topological Discovery

If a genuine application of the interventionist framework is not possible nor desirable, this does not mean that the causal structure and causal behavior of NFRs and NFNs cannot be described by other methods. In the study of nonlinear dynamical systems, qualitative methods are frequently used to describe the causal behavior of the system, partially because analytic solutions are usually unavailable, but also because these methods are deemed most explanatory.

In these methods, what is of interest is not how one variable affects another variable, but rather how the causal behavior of the system as a whole changes as the parameters of the equations are changed. Thus in studies of the Lorenz system, what is of main interest is how the asymptotic behavior changes as the parameter r changes. This is where interesting and identifiable properties of the causal system emerge.

Sometimes the change in behavior in response to change in r is continuous; sometimes the behavior radically changes as r reaches certain critical values (Sparrow 1982).

For instance, for $0 < r < 1$ the origin is stable and globally attracting; at $r = 1$ there is a bifurcation so that for $r > 1$ the origin becomes an unstable equilibrium and two other attracting stable equilibrium points appear. For $r > r_H = \frac{\sigma(\sigma+b+4)}{\sigma-b-1}$ these two equilibrium points become unstable (this is known as a Hopf bifurcation), and we observe the appearance of strange attractors.

The behavior of the Lorenz system is very rich and can be given a much more fine-grained characterization than possible here (for instance, at r slightly smaller than r_H, preturbulence can be observed). What is more important to point out is the method by which the causal behavior of the system is mapped. Two important methods are bifurcation diagrams, which map the counterfactual relation between the parameter and asymptotic behavior, and return maps (or Poincaré maps), which map the intersection of the system's trajectory with a given plane (Figure 3 is a return map).

In bifurcation diagrams and return maps, topological properties are of interest (e.g., the number of attractor states, the shape and size of the attractor states), but not how one individual variable affects another variable in isolation of the whole. The lesson is that nonlinear dynamics theorists do not seek to understand the causal structure of complex nonlinear systems by intervening on individual variables, nor by unrolling the feedback network in time. Instead, they analyze the causal properties of nonlinear systems by more holistic methods which do not seek to isolate a single causal relation between two variables from the causal system as a whole.

Could one not object that the relation between the parameters and the topological properties is itself amenable to interventionist analysis? No doubt interventionist analysis may be applicable to some aspects of a nonlinear feedback system – this does not resolve the problem that it is inapplicable to certain causal relations. In this particular case, it is doubtful the counterfactual relation between parameter value and topological prop-

erties can be called causal. Change in parameter value leads to an instantaneous change in the topological properties, and whether such ‘structural’ relations can be thought of as causal is controversial and subject to debate (Huneman 2010, Skow 2014).

REFERENCES

Angrist, J., and J.-S. Pischke. (2009). Mostly Harmless Econometrics. Princeton: Princeton University Press.

Cartwright, N. (2001). Modularity: It Can – and Generally Does – Fail. In: Maria Carla Galavotti, Patrick Suppes and Domenico Costantini (eds.) Stochastic Causality, CLSI Publications.

—. (2002). Against Modularity, the Causal Markov Condition, and any Link between the Two: Comments on Hausman and Woodward. British Journal for the Philosophy of Science, 53: 411-53.

—. (2004). Causation: One Word, Many Things. Philosophy of Science, 71: 805-819.

—. (2007). Hunting Causes and Using Them: Approaches in Philosophy and Economics. Cambridge, UK: Cambridge University Press.

Clarke, B., B. Leuridan, and J. Williamson. (2014). Modelling mechanisms with causal cycles. Synthese, 191: 1651-1681. DOI: 10.1007/s11229-013-0360-7

Eberhardt, F., and R. Scheines. (2007) Interventions and Causal Inference. Philosophy of Science, 74: 981-995.

Eberhardt, F., B. Glymour, and R. Scheines. (2006). N-1 Experiments Suffice to Determine the Causal Relations Among N Variables. In: Dawn E. Holmes, Lakhmi C. Jain (Eds.), Innovations in machine learning, Berlin: Springer, pp. 97-112.

Gebharter, A., and G. Schurz. (2016). A Modeling Approach for Mechanisms Featuring Causal Cycles. Philosophy of Science, 83: 934-945.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in Bayesian networks. Networks, 20: 507-534.

Guastello, S.J. (2013). Chaos, Catastrophe, and Human Affairs: Applications of Nonlinear Dynamics to Work, Organizations, and Social Evolution. New York: Psychology Press.

Gueégan, D. (2009). Chaos in Economics and Finance. Annual Review in Control, 33: 89-93.

Hausman, D., and J. Woodward. (2004). Modularity and the Causal Markov Condition: A Restatement. British Journal for the Philosophy of Science, 55: 147-61.

Hoyer, P. O., D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. (2009). Nonlinear causal discovery with additive noise models. Advances in neural information processing systems: 689-696.

Huneman, P. (2010). Topological explanations and robustness in biological sciences. Synthese, 17: 213-245.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., and Jarvisalo, M. (2013). Discovering cyclic causal models with latent variables: A general SAT-based procedure. In: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence.

Kaiser, M.I. (2016). On the Limits of Causal Modeling: Spatially-Structurally Complex Biological Phenomena. Philosophy of Science, 83: 921-933.

Kay, S. M. (1993). Fundamentals of statistical signal processing. Upper Saddle River, NJ: Prentice Hall PTR.

Koster, J.T.A. (1996). Markov Properties of Nonrecursive Causal Models. The Annals of Statistics, 24: 2148-2177.

Kuorikoski, J. (2012). Mechanisms, Modularity and Constitutive Explanation. Erkenntnis, 77:361–380. DOI 10.1007/s10670-012-9389-0

Lacerda, G., P. Spirtes, J. Ramsey, and P. O. Hoyer. (2008). Discovering cyclic causal models by independent components analysis. Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, Helsinki.

Mitchell, S. (2008). Exporting Causal Knowledge in Evolutionary and Developmental Biology. Philosophy of Science.

—. (2009). Unsimple Truths: Science, Complexity, and Policy. Chicago: University of Chicago Press.

Neal, R. (2000). On deducing conditional independence from d-separation in causal graphs with feedback: The uniqueness condition is not sufficient. Journal of Artificial Intelligence Research, 12: 87-91.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd edition). Cambridge: Cambridge University Press.

Pearl, J., and R. Dechter. (1996). Identifying Independencies in Causal Graphs with Feedback. Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers.

Reutlinger, A. (2012). Getting rid of interventions. Studies in History and Philosophy of Biological and Biomedical Sciences, 43: 787-795.

Richardson, T. and Spirtes, P. (1999). Automated Discovery of Linear Feedback Models. In: C. Glymour and G. Cooper (Eds.), Computation, Causation, and Discovery, Cambridge, MA: MIT Press, pp. 253-304.

Robinson, J. C. (2004). An introduction to ordinary differential equations. Cambridge, UK: Cambridge University Press.

Skow, B. (2014). Are There Non-Causal Explanations (of Particular Events) British Journal for the Philosophy of Science, 65: 445-467.

Sparrow, C. (1982). The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. New York: Springer-Verlag.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers.

Steel, D. (2006). Comment On Hausman and Woodward on the Causal Markov Condition. British Journal for the Philosophy of Science, 57: 219-231.

Triantafillou, S., and I. Tsamardinos. (2015). Constraint-based causal discovery from multiple interventions over overlapping variable sets. Journal of Machine Learning Research, 16: 2147-2205.

Viswanath, D. (2004). The fractal property of the Lorenz attractor. Physica D, 190: 115-128.

Weber, M. (2016). On the Incompatibility of Dynamical Biological Mechanisms and Causal Graphs. Philosophy of Science, 83: 959-971.

Woodward, J. (1999). Causal Interpretation in Systems of Equations. Synthese, 121: 199-247.

—. (2003). Making Things Happen: A Theory of Causal Explanation, Oxford: Oxford University Press.

—. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. Biology and Philosophy. DOI 10.1007/s10539-010-9200-z

—. (2015). Methodology, ontology, and interventionism. Synthese. DOI 10.1007/s11229-014-0479-1