Synthese (2018) 195:2295–2312
https://doi.org/10.1007/s11229-017-1341-z

Intervening on structure

Daniel Malinsky[1]

Received: 1 June 2016 / Accepted: 15 February 2017 / Published online: 27 February 2017

Abstract Some explanations appeal to facts about the causal structure of a system in order to shed light on a particular phenomenon; these are explanations which do more than cite the causes X and Y of some state-of-affairs Z, but rather appeal to "macro-level" causal features—for example the fact that A causes B as well as C, or perhaps that D is a strong inhibitor of E—in order to explain Z. Appeals to these kinds of "macro-level" causal features appear in a wide variety of social scientific and biological research; statements about features such as "patriarchy," "healthcare infrastructure," and "functioning DNA repair mechanism," for instance, can be understood as claims about what would be different (with respect to some target phenomenon) in a system with a different causal structure. I suggest interpreting counterfactual questions involving structural features as questions about alternative parameter settings of causal models, and propose an extension of the usual interventionist framework for causal explanation which enables scientists to explore the consequences of interventions on "macro-level" structure.

Keywords Causation · Intervention · Counterfactuals · Explanation

1 Introduction

The existence of certain causal dependencies seems to be explanatorily relevant in a variety of contexts, especially in the social and biological sciences. For example, the fact that an economic system is a "free market" and not a "command economy" may play an important role in explaining why certain macroeconomic variables are

✉ Daniel Malinsky
malinsky@cmu.edu

[1] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15206, USA

distributed in a particular way (Hoover 2001, p. 45). A command economy is one in which economic production (among other things) is controlled by the government, so the number of goods produced and resources employed have particular causal dependencies, as decided by the central authority. That is, a command economy will have a particular causal structure in contrast to a free market economy which will have a different causal structure, determined only by unregulated market forces. Compare this story with an explanation in cancer science due to Casini et al. (2011). Casini et al. appeal to the structure of a cell's "DNA repair mechanism" (how various cell subsystems and components respond to DNA damage) in explaining the cancer's progress and responses to certain treatments. When the repair mechanism functions correctly, i.e., when the causal structure has the connections expected of a healthy cell, then one can expect certain behavior. But malfunctioning cells have a different mechanism for repair—a different structure—and this is crucial for understanding cell survival in biological systems affected by cancer. What these two examples from macroeconomics and cell biology share is that they both appeal to features of causal structure (not events) to explain some phenomenon of interest.

What "structure" refers to here includes "what causes what" and also other characteristics of causal relations under study, such as the magnitudes and signs of the various causal connections in a system. The fact that a system has certain causal features, e.g., that A causes B and not the other way around, or that C inhibits D and doesn't promote it (the causal effect is negative not positive), may be of interest for a number of reasons. In light of such facts, one may pose a class of counterfactual questions: for example, what would be different if, contrary to fact, B caused A or if C promoted D? Or what if A had only a "weak" influence on B instead of a "strong" one? Woodward (2003) calls such "what-if-things-had-been-different" questions *w-questions*. In this paper I propose a way of interpreting w-questions when they concern structural features, which extends the usual notion of an intervention. In other words, I propose an interventionist semantics for counterfactuals about causal structure. Further, I suggest techniques for learning the answers to these counterfactual questions; I consider how scientists may use typical data and models to learn about interventions on structural features. The intuitive but often opaque language of changes to structural features, sometimes expressed as the possibility of "modifications of the arrows" in a causal graph (Morgan and Winship 2012), can be understood and explored systematically without abandoning the concepts and tools already employed in causal modeling.

Formally speaking, I will identify structural features with sets of causal parameters (and functions) in structural equation models. I will suggest we interpret counterfactuals about structural features as claims about alternative parameter settings in these causal models, and consequently we can explore the truth-values of these counterfactual claims with a combination of (machine-implementable) data manipulation, simulation, and statistical estimation.

This project has connections to two important questions in the philosophical literature on explanation. There is an interest among philosophers of science in the possibility of non-causal explanation (see, e.g., Strevens 2008, pp. 177–180; Lange 2016). Are all good explanations in empirical science ones which cite relevant causal facts, and thus subject to the constraints or limitations of our "best" theories of causation? Separately, there is a longstanding controversy over the place of structural features in social science (e.g., Jackson and Pettit 1992; List and Spiekermann 2013; Haslanger 2016). Structural features have been understood in various ways, and in these discussions researchers typically inquire into the explanatory role of (e.g.) interpersonal relations or facts about social organization in contrast with facts about individual actors. One can raise the same question for biological systems, and indeed my understanding of structural features should apply to both social and biological science. On my formal characterization of structural features and counterfactuals about them, claims about structure can figure in causal explanations. Note that though I will not defend this claim here, I endorse Woodward's broad notion of causal explanation: "roughly, any explanation that proceeds by showing how an outcome depends (where the dependence in question is not logical or conceptual) on other variables or factors counts as causal," (Woodward 2003, p. 6). A consequence of my view presented below is that any purportedly non-causal explanations must be non-causal by virtue of something other than the fact that they cite structural features. Moreover, though I will not argue for the superiority of explanations which cite either "micro" or "macro" features (i.e., I do not contribute to the debate over "reductionism" in social science explanations), on my proposal structural features can at least sometimes be fruitfully identified with relations among causal variables and thus some important structural explanations can occur on the same "level" as familiar interventionist explanations.

First, I will elaborate on the issue with some examples and explain why the existing literature on causal manipulation does not always suffice to make sense of claims about structural features of scientific interest. I assume a broadly interventionist view of causation along the lines of Woodward (2003) and Pearl (2009), although I will return to the topic of causation in general in the last section.

2 Manipulation and macro-level structural features

It is common in several areas of social science—including sociology, macroeconomics, social epidemiology, and others—to investigate the relationship between social outcomes of interest (e.g., wealth, health, or political stability) and macro-level structural features such as "patriarchy," "globalization," and "healthcare infrastructure" (e.g., Witz 1990; Rodrik 2008; Bhargava et al. 2005). Moreover, concepts like patriarchy, capitalism, globalization, and systemic racism appear in various areas of social philosophy. Concepts such as these are difficult to define and operationalize across contexts; different kinds of studies may call for different understandings of what patriarchy consists in, for example. Yet these ideas figure in explanations, and in particular are often relevant to inferences concerning future policies. The connections between these abstract concepts and concrete policy interventions often remains obscure. What does it mean to intervene on patriarchy? How can we ever learn about how labor conditions might be different in an economic system which is not subsumed under global capitalism? Is it possible to intervene on a society's healthcare infrastructure all at once? These issues are practically important but difficult to answer at any level of generality in the interventionist framework because these diffuse or distributed factors are rarely captured by a single variable in an empirical study, and, when they are explicitly represented as variables, these are probably not features of society which one can manipulate independently of the other variables in the system (Bright et al. 2016, p. 76). Interventionism explicates the semantics of causal claims by reference to possible or hypothetical changes to variables. In order for a variable X to be considered a cause of Y there must be some way to change X which satisfies the conditions laid down in formal theories of causal manipulation (Woodward 2003, pp. 94–99; Pearl 2009, pp. 68–72). Roughly speaking, the intervention on X must not affect Y (if at all) except via X, i.e., confounding dependencies between the intervention and other causes of Y are not allowed. Causal claims about social features including patriarchy and capitalism are confusing because it does not seem like such interventions are possible, even in principle. Prima facie, global capitalism seems to be a cause of the distribution of resources among nations but it is unlikely that any intervention which "ends global capitalism" will not also involve redistribution of resources, in violation of the conditions on an intervention. One option for responding to such considerations is to deny that these concepts are either cogent or causal, or that they can figure in causal explanations (cf. Steel 2006). I propose an alternative: we can interpret (in certain contexts) macro-level social features as facts about the causal structure of society, and we can use the resources available in causal modeling to investigate what might be different under alternative social structures. More generally, I will use the term *structural features* to refer to "facts about the causal structure" of a system, and I will formally identify structural features with collections of structural parameters (and possibly functions) in a causal model. Interventions can be understood as changes in the structural parameters (and/or possibly functional relationships).

Thinking about the consequences of different parameter settings is common practice in certain areas of science. Consider a simple example from macroeconomics. Hoover (2001, p. 48) discusses an economic model relating money supply, interest rates, income, and some other measurable variables. He provides a hypothetical model for the system under study (a set of linear equations, illustrated with a graph in Fig. 1), and notes that the model would be different in theoretically interesting ways if the central bank were to enact a policy to make interest rates constant. The bank can accomplish this by manipulating the causal relationship between income and money supply, so that the influence of income on money exactly counterbalances the influence of money on interest rates.[1] The important point here is that an economic regulation amounts to manipulating parameters in the model, which could mean setting some parameter equal to zero or to some particular non-zero value which balances other influences, making interest rates constant and thus probabilistically independent of money supply. The bank does not intervene directly on interest rates, nor does it arbitrarily increase

[1] This relates to a point raised by Cartwright (1999, pp. 16–17) and others: policies may produce violations of faithfulness (Spirtes et al. 2000). That is because policies which engineer structural parameters on converging paths to exactly cancel make variables appear statistically independent even when they are in fact causally dependent. This can pose problems for applying causal search algorithms that rely on faithfulness as an assumption.

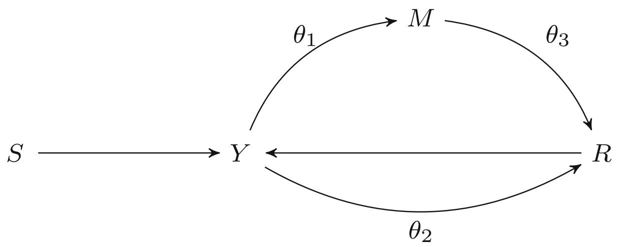

Fig. 1 A graphical illustration of Hoover's (2001, p. 48) example. *M* is money supply, *R* is the interest rate, *Y* is income, and *S* represents "shocks." The notation is changed somewhat. He considers an intervention by the central bank which sets $\theta_1 = -\theta_2/\theta_3$

or decrease the money supply. Instead, as a matter of legal regulation, it changes the economic structure by manipulating the strength of one variable's causal influence on another.[2]

Causal parameters can vary for natural reasons too. For example, plant biologists Bishop and Schemske (1998) study populations of *Lupinus lepidus* (a kind of flowering legume) which colonize heterogeneous environments before and after a volcanic eruption. Their study, which makes use of structural equation modeling (a causal modeling framework discussed in the next section), considers the causal relationships among several plant attributes including total flower number, flowering duration, mean date of inflorescence production, and percentage of flowers damaged. They estimate structural coefficients for multiple plant populations that vary in environmental background and find that the causal dependencies between measured variables differ in different populations. For example, the effect of mean date of inflorescence production on percentage of flowers damaged varies among different environments. The upshot here is that interesting differences in structural features can be produced by natural differences in background ecology. Nobody enacted a policy to regulate structural features in the different plant populations, but the hetergeneous background ecologies are analagous to different policy regimes in the social world.

The macroeconomic example is relatively straightforward in the sense that the structural feature under consideration is just a fact about one causal parameter: the influence of income on money supply. The plant biology example is more complicated because researchers actually consider changes in multiple parameters. Similarly, macro-level features of social structure such as patriarchy and health infrastructure are perhaps most fruitfully interpreted as claims about several structural parameters. A patriarchal social system might be one in which gender is a cause of various features of social life and measures of status: gender is a cause of income, a cause of occupational prestige, a cause of domestic decision-making power, etc. Kaufman (2014) claims that

[2] Note that Hoover would object to my characterization here, because we use different terminology, viz., different definitions of "parameter," "variable," and "structure." I will return to Hoover's alternative view later in the paper, after my own view has been more formally spelled out. In this discussion, I use "variable" and "parameter" as is common in classical statistics and the literature on causal modeling (e.g., Pearl, Woodward, Spirtes et al.): variables are measured quantities and parameters are population-level quantities inferred or estimated from the data.

epidemiologists have a roughly similar notion about racism: to reduce or eliminate racism, for Kaufman, is to eliminate the causal connection between race and socio-economic status (and presumably other measures of well-being). In the context of a particular structural model, a nation's health infrastructure might be constituted by a collection of facts including how an individual's distance from the capital city affects their access to medicine, the fact that wealth strongly influences hospital quality, the fact an individual's frequency of doctor visits depends on what kind of health insurance they have, and so on. In other words, a claim about the healthcare infrastructure is a claim about the causal structure of a society—which causal connections exist and what their strengths are, or which social attributes (in part) determine health outcomes. To answer a w-question about alternative healthcare infrastructures is to ask about how things might turn out in a society where various causal connections are different.

3 Structural features in structural equation models

Many social and biological systems are represented by structural equation models (SEMs). I'll begin formally explicating interventions on structure with reference to a simple class of such models, and then generalize by relaxing some simplifying assumptions.

Consider a single, linear structural equation with 3 measured variables and additive noise:

$$Y = \theta_1 X_1 + \theta_2 X_2 + \varepsilon_Y \tag{1}$$

X_1 and X_2 are (potential) direct causes of Y and the exogenous error variable ε_Y represents the combined influence of all the other causes of Y which are not explicitly represented in the model (Woodward 1999). Implicitly, X_1 and X_2 are also exogeneous, i.e., $X_1 = \varepsilon_{X_1}$ and $X_2 = \varepsilon_{X_2}$ though by convention this is not always written down. The parameters θ_1 and θ_2 represent the strength of the causal connections between X_1 and Y and X_2 and Y respectively; they can be zero (which means no causal influence), positive, or negative. The value of θ_1 indicates how much Y will change given a unit change in X_1.[3] Say in reality $\theta_1 = 0.4$ and $\theta_2 = -1.3$. The vector $(\theta_1, \theta_2) = (0.4, -1.3)$ summarizes a lot about the causal structure of the system: it tells us what causes what (which parameters are zero) and what the strengths of the causal connections are. In the context of linear SEMs, my proposal amounts to identifying interventions on structure with settings of parameters (adapting Pearl's *do* notation): $do(\theta_1, \theta_2 = \tilde{\theta}_1, \tilde{\theta}_2)$, where the tilde indicates new, specific values for individual parameters. One can intervene to change $(\theta_1, \theta_2) = (0.4, -1.3)$ to $(0, 0.87)$ or $(0.4, -0.2)$ or some other vector. This may be the result of some real or imagined policy, like in Hoover's interest rate example.

[3] If we leave the world of SEMs and instead model a system with a parameterized, discrete-valued Bayes net, then the variables can each take on a finite number of "states" and parameters are just state transition probabilities.

More generally, an SEM may consist of a number of structural equations, which may or may not be linear. For example:

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= f_2(X_1; \alpha) + \varepsilon_2 \\ X_3 &= f_3(X_1, X_2; \beta, \gamma) + \varepsilon_3. \end{aligned} \tag{2}$$

Each variable X_i is related to its direct causes by some smooth function, f_i. This model involves no feedback (the graphical representation will be *acyclic*). Here X_1 has no causes except the error term. Each function is parameterized by some set of parameters. For example, f_3 may be $\beta\exp(-X_1/\gamma)+\sin(\gamma X_2)$. An intervention on the structure of such a system is a setting $do(f_2, f_3, \alpha, \beta, \gamma = \tilde{f_2}, \tilde{f_3}, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$. That is, we replace the functions with (possibly) new functions and the parameters with (possibly) new values. There is already some work which explores interventions on functions. Mooij and Heskes (2013) investigate changes to a system which amount to changing the function relating Y to its direct causes, i.e., replacing $Y = f(X_1, X_2, \varepsilon)$ with $Y = \tilde{f}(X_1, X_2, \varepsilon)$, where X_1 and X_2 are direct causes of Y.[4] The scientific application which Mooij and Heskes have in mind is the study of protein-signaling networks in cellular biology. Changing the functional relationship between two variables is meant to model the kind of experiments which—in Mooij and Heskes' terminology—change the "activity" (not "abundance") of compounds in the cell.

Most abstractly, an intervention on the structure of an SEM is a setting of the functions and parameters which characterize the model to a new set of functions and a new set of parameter values: $do(\{f_i\}, \{\theta_i\} = \{\tilde{f_i}\}, \{\tilde{\theta_i}\})$. There is one function for each variable that appears in the model, i.e., the function which relates the ith variable to every other variable.[5] θ_i is the corresponding vector of parameters for function f_i.

Interventions on structure defined in this way should be familiar to economists (e.g., Cooley and LeRoy 1985). The definition here is also a general version of what Tian and Pearl (2001) call a "mechanism change." In that paper, Tian and Pearl are working in the framework of directed acyclic graphs (DAGs), and they are principally concerned with learning causal structure from data when mechanism changes—changes in causal parameters—are estimated to have taken place. Note that there are connections between this work and the econometric literature on "structural breaks" (e.g., Hoover and Sheffrin 1992; Perron 2006), though I will not elaborate on the connection here. Steel (2006) discusses interventions which he calls "structure-altering interventions"; these are formally different from the interventions I consider because Steel's interventions directly change the distributions of measured variables as well as the causal connections among them (thus they are non-ideal interventions on variables in Woodward's (2003) terminology). Interventions on structure as I have defined them

[4] ε is inside the function because the authors do not assume that the causal mechanism is necessarily additive in the noise variable. They also allow for causal feedback; more on that below.

[5] Note that some variables may not be causally related and some may have no causes at all among the measured variables, only error terms. This is all easy to accomodate with functions defined accordingly: functions can be independent of some (or all) of their arguments, and so we can assume every variable is an argument of every function to be maximally general.

only directly change causal connections, though they can indirectly affect "downstream" distributions of measured variables.[6]

So far, I've only proposed a formal definition of an "intervention on structure" and in the next section I will discuss how we might infer the consequences of such an intervention. My definition is analagous to the standard notion of intervention on variables found in Pearl (2009), Woodward (2003), and Spirtes et al. (2000), where an intervention is a change to some feature of the model but which leaves most of the rest of the model intact. In the Pearl/Woodward/SGS case, the distribution of some variable changes, and the variable becomes disconnected from its causes; the rest of the model is not directly affected. In my case, the value of some parameter (or function) changes, and since causal parameters are not "caused" by other parameters or variables in the model, nothing else is changed by the setting; the rest of the model is not directly affected. My definition is motivated by certain common counterfactual questions, and so it might help to give some examples.

Consider again model (1), with parameter values $(\theta_1, \theta_2) = (0.4, -1.3)$. Typically to fill out the model we also assume a probability distribution for the exogenous variables as part of the model, e.g., that $\varepsilon_{X_1}, \varepsilon_{X_2}, \varepsilon_Y \sim N(0, 1)$. We might ask: "what would be the distribution of Y if X_1 had a stronger effect on Y, say twice as strong?" On my construal, we can make this precise: "what would be the distribution of Y given $do(\theta_1 = 0.8)$?" In order to answer this question, we would need to calculate the distribution of Y in the model:

$$Y = 0.8X_1 - 1.3X_2 + \varepsilon_Y,$$

with $\varepsilon_{X_1}, \varepsilon_{X_2}, \varepsilon_Y \sim N(0, 1)$ as before. Alternatively we might ask: "what would be the distribution of Y if X_2 did not cause Y, i.e., X_2 had precisely zero effect on Y?" On my construal, this is: "what would be the distribution of Y given $do(\theta_2 = 0)$?" In order to answer this question, we would need to calculate the distribution of Y in the model:

$$Y = 0.4X_1 + 0X_2 + \varepsilon_Y,$$

with $\varepsilon_{X_1}, \varepsilon_{X_2}, \varepsilon_Y \sim N(0, 1)$ again. We might even ask: "what would be the joint distribution over all the variables if the causal relationship between Y and X_1 were flipped?" Things get more complicated here, but on my construal this is asking about the resultant distribution given $do(\theta_1, \theta_3 = 0, 0.4)$ where θ_3 is the causal effect of Y on X_1 (previously implicitly zero). That is, we would need to calculate the joint distribution for the variables in the model:

$$\begin{aligned} Y &= 0X_1 - 1.3X_2 + \varepsilon_Y \\ X_1 &= 0.4Y + \varepsilon_{X_1}, \end{aligned}$$

[6] Steel's focus in (2006) is invariance and explanation across macro/micro descriptions of social systems. He is not concerned with learning the truth-values of structure-altering counterfactual claims, so he does not provide any procedure for doing so. The examples he uses to motivate his definition of structure-altering interventions may be quite plausible, and it would be illuminating to work out a technique for predicting the outcomes of structure-altering interventions to compare with the proposal in the next section.

with $\varepsilon_{X_1}, \varepsilon_{X_2}, \varepsilon_Y \sim N(0, 1)$. This last counterfactual may be a strange one to consider—there may be no plausible policy which corresponds to a change in θ_1 and θ_3 but which keeps the distribution of ε_{X_1} unchanged—but the counterfactual question can be posed precisely. I will return to this issue and the issue of *calculable* interventions on structure in the next section. Note that one may likewise pose counterfactual questions about structural change in possibly non-linear causal models, even models with interaction effects: e.g., "what would be the distribution of Y if the functional dependence of Y on X_1 and X_2 were multiplicative rather than additive?"[7]

Consider the following more concrete hypothetical example. One may model part of the healthcare system in a developing country with an SEM relating infant mortality to mother's age, mother's education, hospital quality, and access to reproductive services. Say this last variable in turn depends on mother's education (maybe certain reproductive services are distributed through schools), and both mother's education and hospital quality depend on distance from an urban center. So we have this (grossly oversimplified) model:[8]

$$\begin{aligned} \text{Rep services} &= \theta_1 \text{Education} + \varepsilon_1 \\ \text{Hospital qual} &= \theta_2 \text{Distance} + \varepsilon_2 \\ \text{Education} &= \theta_3 \text{Distance} + \varepsilon_3 \\ \text{Infant mortality} &= \theta_4 \text{Age} + \theta_5 \text{Education} + \theta_6 \text{Hospital qual} \\ &\quad + \theta_7 \text{Rep services} + \varepsilon_4 \end{aligned} \tag{3}$$

with all errors mutually independent and distributed according to some assumed probability distribution. Say we estimate that

$$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7) = (1.2, -0.8, -0.9, 0.5, -0.4, -1.4, -0.6).$$

An intervention on the healthcare infrastructure might be one which makes access to reproductive services independent of mother's education (by evenly dispersing clinics which provide reproductive services through the region, thus taking schools out of the equation), and makes hospital quality only weakly dependent on distance from the capital city (by allocating more hospital staff and funding to rural hospitals). So we consider the intervention

$$do(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7 = 0, -0.15, -0.9, 0.5, -0.4, -1.4, -0.6).$$

[7] See Bright et al. (2016). Note that some counterfactual questions may be ambiguous, especially in the context of non-linear models. For example, to ask "What would be the distribution of Y if the causal effect of X_1 was twice as strong?" relative to the model $Y = 0.4X_1 - 1.3X_2 + 0.2X_1X_2$, is ambiguous because I could have in mind the coefficient on X_1, the coefficient on the interaction term X_1X_2, or some combination of these. Each option corresponds to a different intervention.

[8] I omit equations for distance and the other exogenous variables from the model. Of course it is quite likely that a better model is not linear and includes many more factors, but this model is for illustrative purposes only. See Bhargava et al. (2005) for an actual empirical study which partially inspired this example.

Now we can ask a precise w-question about the distribution of infant morality under this alternative parameter setting.

This admittedly sketchy example brings out how one can explicate something corresponding to "healthcare infrastructure" in the context of a particular model. Similarly, "patriarchy," "capitalism," and other macro-structural features of a system may be identified with sets of parameters (and/or functions) in some model and we can ask, for example, "what would be different about the distribution of wealth if there were no gendered wage gap (no causal dependence of salary on gender), and if hiring/promotions were gender-blind?" Policy proposals like the legal regulation of wages based on job category or double-blind review in promotions decisions are attempts at manipulating these causal dependencies.

It is worth briefly contrasting my proposal with an alternative formalism developed by Hoover (2001, 2011, 2012, 2013), who also considers interventions on parameters in several works. Hoover defines his terms quite differently than I do here. For example, for Hoover "parameters" are those quantities which can be directly controlled. Thus, all interventions are interventions on parameters; parameters are the loci of interventions, as Hoover defines them. (Variables are controlled only indirectly through control of corresponding parameters.) This leads Hoover to an importantly different definition of structure. To put it briefly, $Y = 0.4X_1 - 1.3X_2 + \varepsilon_Y$ and $Y = 0.8X_1 - 1.3X_2 + \varepsilon_Y$ correspond to the same structure on Hoover's view, because he identifies only differences in causal ordering with differences in structure; on my definitions, the two parameter settings correspond to different structures. On Hoover's formalism, interventions which transform X_1 into a strong cause of Y (where X_1 was previously a weak cause) are not anything special but rather routine. On the other hand, $Y = 0X_1 - 1.3X_2 + \varepsilon_Y$ and $Y = -1.3X_2 + \varepsilon_Y$ are for Hoover two different structures, since X_1 is not a (potential) cause of Y in the second equation; on my view, they are the same structure, since both amount to a setting of zero for the causal effect of X_1 on Y (and they are observationally equivalent). Though there are differences between Hoover's broader causal theory (what he calls the "structural account") and the interventionist theory of causation associated with Pearl and Woodward, I believe one may consistently adopt either framework for a given analysis. My goal in this paper is to show how the interventionist framework, which has been fruitfully applied in many scientific domains but has hitherto been developed with only interventions on variables (e.g., $do(X = x)$) in mind, can be extended to deal with changes to parameters and functions—what I call structural features. Thus the account of causal explanation which derives from the interventionist formalism can be effectively extended to a broader class of phenomena. I hope that the foregoing examples demonstrate that this account is not inappropriate to capture questions of real scientific interest.

When we ask about what a system might look like with a different causal structure, we are asking about the outcome(s) of an intervention on structural features, where only those variables which are causally "downstream" of the manipulated parameters (or functions) are affected. Should we be wary of interventions on structure? Are they somehow less legitimate, perhaps in a metaphysical sense, than more familiar interventions on variables? I think any such worries should be relieved by the following observation: any intervention on structure can be redescribed as an intervention on a certain kind of variable.[9] In particular, an intervention on structure can be thought of as intervention on some hidden "policy variable" which interacts with other variables in the system. Consider again Eq. (1) and the the intervention $do(\theta_1, \theta_2 = \tilde{\theta}_1, \theta_2)$. This can be redescribed as $Y = \theta_1 X_1 + \theta_2 X_2 + (\tilde{\theta}_1 - \theta_1) X_1 P + \varepsilon$, where P is an indicator variable which is 1 when the policy is "on" and 0 when it is "off". Setting $P = 1$ means that the relationship between Y and X_1 is determined by the new parameter, $\tilde{\theta}_1$. In the social sciences, a variable such as P which affects the relationship between X_1 and Y is called a "moderator" or "pure moderator" (MacKinnon 2008, ch. 10). Thus, interventions on structural features are just equivalent to interventions on (hidden) moderator variables. The same can be said in the context of non-linear models. Say $Y = f(X_1; \theta) + \varepsilon$ and we're interested in $do(f, \theta = \tilde{f}, \tilde{\theta})$. Then we can write $Y = f(X_1; \theta) + (\tilde{f}(X_1; \tilde{\theta}) - f(X_1; \theta))P + \varepsilon$. The specified intervention on structure is equivalent to an intervention which sets the policy variable to 1, and more complicated cases can be captured by the inclusion of multiple policy variables. Interventions on structure do not require any radical departure from interventionist metaphysics, but they can aid in our understanding of explanations which appeal to structural features and not events.

Note that in the context of a given study, e.g. an experiment or analysis which establishes some structural equation(s) from data, researchers will typically have no useful information about P; the "policy" is "off" in the population under study and we are interested precisely in the counterfactual population in which P is "on."[10] It is natural to ask whether such a policy variable actually exists (or exists "in principle"), i.e., whether there is a feasible policy which interacts with the other variables in the specified way and which someone can enact. I will return to this existential question in Sect. 5. The expanded system of equations with added policy variables is only a mathematical device, a redescription of the system which makes salient the fact that changes to parameters are not formally different from changes to variables.

One could model all these hidden policy variables directly, although this would be cumbersome and not much would be gained by doing so. For a range of interventions under consideration, one could write down an extended model with multiple policy variables which are indicators, with all the right interactions to produce the intended parameter setting when the policy is "on." There might be benefits to such an exercise, but the important point remains that counterfactuals about structure are just statements about alternative parameter settings of causal models. The key question is: how can we know the outcomes of such interventions from data, when we have not actually observed the system with the policy "on"?

[9] I am indebted to Greg Gandenberger for raising this point.

[10] More technically, I am not assuming that researchers will have a joint probability distribution over all the variables in the expanded model which includes P, nor even that such a joint probability distribution would be well-defined in complex, non-recursive cases. My only claim is that one can always write down an expanded system of equations with moderator variables which is equivalent to the actual observed system when the policy is "off" and which is describes the counterfactual system when the policy is "on."

4 Learning about the outcomes of interventions on structure

In the previous section I explicated counterfactual questions about structure in terms of interventions on structural features. It is a separate issue which such counterfactual questions have well-defined and calculable answers. In a structural equation model, causal parameters are assumed to be independent, i.e., variation free. This means that θ_1 can take on any value in its range of admissible values independent of the value of θ_2, and likewise for the other parameters. Zhang et al. (2015) discuss more formally the conditions under which parameters are variation free, and relate this assumption to other causal and statistical modeling assumptions. Here I follow common practice in structural equation modeling (as well as causal modeling with DAGs or other representations) in stipulating that the identified parameters in the SEM are variation free, and that the ranges of admissible values for the parameters are set by background knowledge of the domain.[11] In a (possibly non-linear) SEM with no feedback, the variables have a well-defined joint probability distribution for all values of the parameters (or functional dependencies). In fact, if one knows the probability distributions of the exogenous errors, than one can derive analytically the joint probability distribution of all the variables, or any marginal distribution, for any parameter setting.

To emphasize this point, consider the Manipulation Theorem of Spirtes et al. (2000, p. 51). Spirtes et al. consider causal models represented by DAGs which satisfy the Causal Markov Condition. They define interventions on variables in much the same way as Pearl (2009) and Woodward (2003): "ideal" or "surgical" settings of variables to fixed values. The Manipulation Theorem says that if a causal graphical model satisfies the Causal Markov Condition, then the outcome (implied joint probability distribution) of such an intervention is well-defined, and can be calculated using a particular formula from the observational distribution. Similarly, since SEMs with no feedback have well-defined and calculable implied joint probability distributions for all parameter values, interventions on parameters (just settings of the parameters to particular values) imply well-defined and calculable probability distributions.

Unfortunately, the story is not so simple for models which exhibit causal feedback. Systems with feedback can be represented by cyclic causal graphs or non-recursive SEMs (Spirtes 1995; Richardson 1996; Mooij et al. 2011). It is typically assumed that the system is measured at some kind of equilibrium. Then, the outcome of an intervention on the system is understood as a change which (potentially) brings the system to a new equilibrium, i.e., a new joint probability distribution over the variable set. Feedback mechanisms impose constraints on the functions and structural parameters which can be fruitfully considered as candidate "policies." Roughly speaking, a model has a well-defined equilibrium distribution only so long as the functions do not "blow up," i.e., create an unstable feedback process. Consequently, some interventions on sets of structural parameters or functions make counterfactual prediction impossible. In the linear case, there is a well-defined equilibrium distri-

[11] See Woodward (1999) comments on invariance, and also Cartwright (2003). They have in mind invariance with respect to interventions on variables, but many of the same considerations carry over to invariance with respect to interventions on structural features.

bution only when the structural coefficients involved in the feedback loop satisfy a mathematical constraint (see Fisher 1970).[12] That means that only those interventions on structural parameters that satisfy this mathematical constraint will lead to stable outcomes. There are analagous mathematical constraints in non-linear models; Mooij et al. (2011) derive formal conditions for non-linear, non-recursive models with Gaussian errors and only two variables. More general formal conditions for systems with many variables can be investigated, but the upshot will be the same. Only those interventions on structural relationships which lead to equilibrium can really be explored in non-recursive structural equation models. The same is true for structural models of dynamic systems (i.e., models of stochastic processes). Parameters in such dynamic models must satisfy certain mathematical constraints in order to be well-behaved in a statistical sense; otherwise the system does not have a stable distribution. Thus, in considering candidate interventions on structure in models with feedback or dynamical processes, we can only investigate counterfactual settings which respect these mathematical constraints, or else we are setting ourselves up to consider counterfactual predictions for an unpredictable system. Moreover, even in SEMs with no feedback, we have to be careful that the proposed intervention does not create a feedback loop, or else the same mathematical constraints apply.

For the remainder of the paper, assume the model is a recursive SEM (no feedback), and consider only interventions which create no feedback loops. Though it is possible to calculate the implied probability distribution of an intervention on structure analytically when the distributions of exogenous variables are known, sometimes the distributions are not known, and we would like to estimate the counterfactual distribution from available data. For example, consider again the simple model (1):

$$Y = \theta_1 X_1 + \theta_2 X_2 + \varepsilon_Y$$

Assume that one has access to enough data on the measured variables X_1, X_2, and Y to establish Eq. (1) reliably, with parameter estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Estimate the empirical distribution of ε_Y by taking the residuals, i.e., $Y - \hat{\theta}_1 X_1 - \hat{\theta}_2 X_2$. Now, we can ask the following w-question: what would be the distribution of Y if $\theta_1 = \tilde{\theta}_1$ instead of $\hat{\theta}_1$? Well, the influence of X_2 and ε_Y would not change, by assumption. So one could re-use the data to estimate the distribution of Y in the following way. Take the measurements on X_2, multiply them by $\hat{\theta}_2$. Add these to the measurements on X_1, multiplied by the new parameter value $\tilde{\theta}_1$. Add to each data point in this new vector a number drawn from the error distribution. Now the researcher has a new set of values for Y. Using any number of density estimation methods (e.g., kernel based methods, histograms), the researcher can construct a counterfactual empirical probability distribution for Y. Alternatively, one may estimate the mean of Y if that is the only quantity of interest. The logic behind this informal algorithm is just that a change to θ_1 implies a change to the data downstream from θ_1. Using the data and standard statistical techniques, the posed w-question can be answered.

[12] The eigenvalues of the coefficient matrix of the non-recursive structural equations must have modulus less than or equal to 1.

Consider a slightly more complicated example, where the model consists of a system of linear structural equations:

$$\begin{aligned} X_2 &= \theta_{2,6}X_6 + \varepsilon_2 \\ X_1 &= \theta_{1,3}X_3 + \theta_{1,5}X_5 + \varepsilon_1 \\ X_4 &= \theta_{4,1}X_1 + \theta_{4,2}X_2 + \varepsilon_4 \end{aligned} \tag{4}$$

This system is just like the first one: it is linear, with no feedback, and assume the error variables are all jointly independent, meaning that there are no unmeasured common causes in the system (the model is "causally sufficient" in the terminology of Spirtes et al. 2000). X_1 is a direct cause of X_4, but also has its own direct causes, X_3 and X_5. Imagine an intervention $do(\theta_{1,3}, \theta_{1,5}, \theta_{2,6}, \theta_{4,1}, \theta_{4,2} = \theta_{1,3}, \tilde{\theta}_{1,5}, \theta_{2,6}, \tilde{\theta}_{4,1}, \theta_{4,2})$. The only parameters which change are $\theta_{1,5}$ and $\theta_{4,1}$. To determine the distribution of X_4 after this manipulation we proceed as before, but with an extra step which reflects the causal order: to know the distribution of X_4 we need to know X_1, which depends on X_5. So, take the data on X_5 multiplied by the new parameter $\tilde{\theta}_{1,5}$, and add it to $\hat{\theta}_{1,3}$ times the measurements on X_3. Add a random deviate to each data point from the empirical distribution of ε_1. The result is a new set of data points for X_1. Use these data points, multiplied by $\tilde{\theta}_{4,1}$ and added to $\hat{\theta}_{4,2}$ times X_2 plus points from the empirical distribution of ε_4. The result is a new set of data points for X_4. One can go on to estimate the density of X_4 with their preferred density estimation technique.

To proceed in this manner for the hypothetical healthcare infrastructure example from Sect. 3, we would consider the distribution of reproductive services after clinics were evenly dispersed (i.e., $do(\theta_1 = 0)$), and the distribution of hospital quality after hospital staff and funding was re-allocated to rural hospitals (weaking the influence of distance by setting $do(\theta_2 = -0.15)$). With these new distributions we estimate the counterfactual distribution of infant mortality using the last equation.

The reasoning above is quite transparent, and it is easy to generalize for linear structural equation models with no feedback. Given a model, some data, and a new vector of parameters one may calculate the post-intervention distribution of some variable Y by following a sequence of data manipulations based on the model. First identify the causes of Y and their causes and so on. Begin by recalculating the data points for each variable causally upstream from Y using the appropriate combination of weighted data vectors and error variables.[13] Move down in the causal order toward Y until arriving at a new set of data points for Y, and then estimate the density. This set of steps would be easy to implement on a computer. The performance of the procedure, including the estimation techniques involved (at varying sample sizes) and the consequences of small errors, can be investigated by simulation.

[13] If the quantity of interest is only the post-manipulation mean of Y, and not its entire density function, then it may be prudent to simply add the mean of the relevant error variable at each step. The errors are assumed to be Gaussian with mean zero in many typical models, in which case this is trivial. However, more generally if one is interested in the distribution of Y and the error variables are not normally distributed, i.e., if there is important information in the error distribution, then simply adding the mean of ε will not suffice. When the errors are significantly non-Gaussian, then the full distribution of ε is likely relevant. Thanks to Jonathan Livengood for raising this issue.

The same reasoning can be extended to SEMs which are not linear, such as in equation system (2). Non-linearities and non-Gaussian error distributions make the statistical problem progressively more difficult—requiring larger sample sizes, more computational resources, perhaps sacrificing statistical reliability—but the underlying logic is simple. The work involved (estimation and re-estimation of quantities based on the model, simulations to assess reliability or robustness) is no different from the kind of work which already goes on in empirical social and biological sciences. In fact, something approximating my suggestion is already common practice in areas of empirical economics (see Aguirregabiria and Ho 2012).

It is worth emphasizing that on this view, the truth or falsity of a counterfactual about structure is tied to a particular causal model. Fix a causal structure (in this case, an SEM with no feedback) and then there are clear rules for answering w-questions about structural features. Sometimes, a whole class of causal models which share important features (maybe all these models agree that X causes Y which causes Z, but disagree about whether Z causes W) can imply the same verdict on a given w-question because not every causal fact is relevant to a particular counterfactual.[14] In any case, when we interpret the meaning of structural counterfactuals in the manner suggested here, we also have (the beginnings of) an epistemology for learning which counterfactuals are true.

Learning about counterfactuals on structure from observational data is an even harder problem than the usual problem of learning standard interventionist counterfactuals. Though this ought to be thoroughly investigated by simulation, the technique sketched here is bound to leave a substantial amount of uncertainty in the answer to any particular question. Thus, while I have argued that counterfactuals about structural features are well-defined and knowable in principle, in practice available data may rarely resolve counterfactual predictions about changes in “healthcare infrastructure,” “patriarchy,” and the like. Possible problems for future research may include investigating finite-sample performance of counterfactual estimation techniques via simulation (where the truth is known) and inquiring into possible statistical consistency guarantees.

5 Are structural features really causes?

Throughout this discussion I have assumed that social systems and biological systems exhibit causal structure: they can be represented as systems of causally interpreted structural equations, or parameterized causal graphical models. All of Woodward (2003), Pearl (2009) and Spirtes et al. (2000) agree that the causal relata are random variables. For all of these authors, the notion of an intervention plays a central role in the epistemology of causation. None of them consider structural features, as I have defined them, to be causes.

One difference between Woodward and Spirtes et al. is that Woodward defines causation in terms of interventions (A is a cause of B iff *there exists* an intervention on A satisfying certain properties) while Spirtes et al. take what may be called an axiomatic approach where “direct cause” is a primitive: roughly speaking, A directly

[14] Thanks to Conor Mayo-Wilson for discussion on this point.

causes B with respect to some set of variables when $A \rightarrow B$ in a model which satisfies the Causal Markov Condition over that set. From this stipulation and several other conditions they arrive at a calculus of interventions, which specifies the consequences of an intervention, *if there is one* (Glymour 2004; see also Glymour and Glymour 2014). Whether there actually exists an intervention with such-and-such properties does not determine whether A is a cause of B.

My suggestion in this paper is in the spirit of this latter view, even though structural features are not among the causal relata in Spirtes et al. and they do not develop a calculus for interventions on structural features. Whether or not there exists a practically feasible (or possible "in principle") intervention which sets (θ_1, θ_2) to $(\tilde{\theta}_1, \tilde{\theta}_2)$ and leaves the causal model otherwise adequate is an important question, but it is a question which can only be answered by careful attention to the domain of empirical application. My approach to the w-questions which are the focus of this paper is epistemological; I suggest a way that scientists might interpret and learn about the consequences of interventions on structural features, whether they are hypothetical or actual. If one endorses the prerequesite that an intervention on θ must be possible for a causal counterfactual about θ to be true (or false), this does not in any serious way conflict with my proposal. It does, however, significantly limit the number of potentially interesting or useful counterfactuals which scientists may add to their body of knowledge. Though some parameters may enjoy the special status of "fundamental constants" in areas of physics, physicists routinely investigate what things might be like with alternative values for these parameters—think, for example, of simulations in cosmology which tell us about how the universe would evolve were the curvature of space different. We probably cannot intervene to make space curved instead of (nearly) flat, but we can surely reason about what that would be like, and appeal to that reasoning in the course of explaining some phenomenon. Counterfactuals about such structural features clearly do play a role in explanations. Thus, an interventionist explication of counterfactuals about structural features places such explanations firmly within the purview of causal explanation. (Recall that I endorse Woodward's broad notion of causal explanation.) One upshot of this view is that when we evaluate the adequacy or success of explanations that appeal to structural features, we can appeal to the same adequacy criteria proposed for causal explanations, keeping in mind the thorny epistemological worries raised in Sect. 4. There is much further work to do in this domain, but I hope that the preceeding discussion has demonstrated that the kind of hurdles involved in establishing counterfactual claims about structural features are just scientific challenges, which we can confront with the same inferential methods employed in other areas of scientific inquiry.

Acknowledgements I would like to thank Liam Kofi Bright, David Danks, Greg Gandenberger, Clark Glymour, Kevin Hoover, Jonathan Livengood, Conor Mayo-Wilson, Jim Woodward, and two anonymous referees for comments on earlier drafts.

References

Aguirregabiria, V., & Ho, C. Y. (2012). A dynamic oligopoly game of the US airline industry: Estimation and policy experiments. *Journal of Econometrics*, *168*(1), 156–173.

Bhargava, A., Chowdhury, S., & Singh, K. K. (2005). Healthcare infrastructure, contraceptive use and infant mortality in Uttar Pradesh, India. *Economics and Human Biology*, *3*(3), 388–404.

Bishop, J. G., & Schemske, D. W. (1998). Variation in flowering phenology and its consequences for lupines colonizing Mount St. Helens. *Ecology*, *79*(2), 534–546.

Bright, L. K., Malinsky, D., & Thompson, M. (2016). Causally interpreting intersectionality theory. *Philosophy of Science*, *83*(1), 60–81.

Cartwright, N. (1999). Causal diversity and the Markov condition. *Synthese*, *121*(1), 3–27.

Cartwright, N. (2003). Two theorems on invariance and causality. *Philosophy of Science*, *70*(1), 203–224.

Casini, L., Illari, P. M., Russo, F., & Williamson, J. (2011). Models for prediction, explanation and control. *Theoria. Revista de Teora, Historia y Fundamentos de la Ciencia*, *26*(1), 5–33.

Cooley, T. F., & LeRoy, S. F. (1985). Atheoretical macroeconometrics: A critique. *Journal of Monetary Economics*, *16*(3), 283–308.

Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica*, *38*(1), 73–92.

Glymour, C. (2004). Review of James Woodward, Making things happen: A theory of causal explanation. *British Journal for the Philosophy of Science*, *55*(4), 779–790.

Glymour, C., & Glymour, M. R. (2014). Commentary: Race and sex are causes. *Epidemiology*, *25*(4), 488–490.

Haslanger, S. (2016). What is a (social) structural explanation? *Philosophical Studies*, *173*(1), 113–130.

Hoover, K. D. (2001). *Causality in macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K. D. (2011). Counterfactuals and causal structure. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 338–360). Oxford: Oxford University Press.

Hoover, K. D. (2012). Causal structure and hierarchies of models. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 778–786.

Hoover, K. D. (2013). Identity, structure, and causal representation in scientific models. In H. Chao, S. Chen, & R. Millstein (Eds.), *Towards the methodological turn in the philosophy of science: mechanism and causality in biology and economics* (pp. 35–60). Dordrecht: Springer.

Hoover, K. D., & Sheffrin, S. M. (1992). Causation, spending, and taxes: Sand in the sandbox or tax collector for the welfare state? *The American Economic Review*, *82*, 225–248.

Jackson, F., & Pettit, P. (1992). Structural explanation in social theory. In D. Charles & K. Lennon (Eds.), *Reduction, explanation and realism* (pp. 97–131). Oxford: Oxford University Press.

Kaufman, J. S. (2014). Commentary: Race: Ritual, regression, and reality. *Epidemiology*, *25*(4), 485–487.

Lange, M. (2016). *Because without cause: Non-causal explanations in science and mathematics*. Oxford: Oxford University Press.

List, C., & Spiekermann, K. (2013). Methodological individualism and holism in political science: A reconciliation. *American Political Science Review*, *107*(4), 629–643.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Routledge.

Mooij, J. M., Janzing, D., Heskes, T., & Schölkopf, B. (2011). On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems* (pp. 639–647).

Mooij, J., & Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the 29th annual conference on uncertainty in artificial intelligence* (pp. 431–439). AUAI Press.

Morgan, S., & Winship, C. (2012). Bringing context and variability back into causal analysis. In H. Kincaid (Ed.), *Oxford handbook of the philosophy of the social sciences* (pp. 319–354). Oxford: Oxford University Press.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge: Cambridge University Press.

Perron, P. (2006). Dealing with structural breaks. *Palgrave Handbook of Econometrics*, *1*, 278–352.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the twelfth international conference on uncertainty in artificial intelligence* (pp. 454–461). Morgan Kaufmann Publishers Inc.

Rodrik, D. (2008). *One economics, many recipes: Globalization, institutions, and economic growth*. Princeton: Princeton University Press.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 491–498). Morgan Kaufmann Publishers Inc.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge: MIT Press.

Steel, D. (2006). Methodological individualism, explanation, and invariance. *Philosophy of the Social Sciences*, *36*(4), 440–463.

Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge: Harvard University Press.

Tian, J., & Pearl, J. (2001). Causal discovery from changes. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 512–521). Morgan Kaufmann Publishers Inc.

Witz, A. (1990). Patriarchy and professions: The gendered politics of occupational closure. *Sociology*, *24*(4), 675–690.

Woodward, J. (1999). Causal interpretation in systems of equations. *Synthese*, *121*(1), 199–247.

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

Zhang, K., Zhang, J., & Schölkopf, B. (2015). Distinguishing cause from effect based on exogeneity. In R. Ramanujam (Ed.), *Proceedings of the fifteenth conference on theoretical aspects of rationality and knowledge, TARK* (pp. 261–271).