Erkenntnis
https://doi.org/10.1007/s10670-021-00463-0

ORIGINAL RESEARCH

Intervening and Letting Go: On the Adequacy of Equilibrium Causal Models

Naftali Weinberger[1]

Received: 18 June 2020 / Accepted: 10 September 2021

Abstract
Causal representations are distinguished from non-causal ones by their ability to predict the results of interventions. This widely-accepted view suggests the following adequacy condition for causal models: a causal model is adequate only if it does not contain variables regarding which it makes systematically false predictions about the results of interventions. Here I argue that this condition should be rejected. For a class of equilibrium systems, there will be two incompatible causal models depending on whether one intervenes upon a certain variable to fix its value, or 'lets go' of the variable and allows it to vary. The latter model will fail to predict the result of interventions on the let-go-of variable. I argue that there is no basis for preferring one of these models to the other, and thus that models failing to predict interventions on particular variables can be just as adequate as those making no such false predictions. This undermines a key argument (Dash in Caveats for causal reasoning with equilibrium models. University of Pittsburgh. PhD thesis, 2003) against relying upon causal models inferred from equilibrium data.

1 Introduction

Dynamic causal models (Iwasaki and Simon 1994; Voortman et al. 2012; Blom et al. 2020) provide graphical tools for representing and inferring the causal relationships in systems that are away from equilibrium. While standard causal modeling methods (Pearl 2009; Spirtes et al. 2000) suffice for systems at equilibrium, dynamic causal models further employ time-derivatives and differential equations to represent the feedback loops by which dynamical systems maintain their equilibrium states. Dynamic causal models would initially appear to provide a generalization of causal models that, while important, could nevertheless be put to the side when studying systems at equilibrium. Yet Dash (2003) argues that, for a class

✉ Naftali Weinberger
naftali.weinberger@gmail.com

[1] Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Ludwigstr. 31, 80539 Munich, Germany

Published online: 11 October 2021

of dynamical systems, the causal models one would infer from sampling their variables only at equilibrium will falsely represent the system's causal relationships, and dynamic causal models enable one to specify which these are. On this picture, rather than being a complement to equilibrium models, dynamic causal models serve as substitute, since for some systems only the dynamic model is adequate. In what follows, I will argue that Dash misdiagnoses the flaw with equilibrium models, and that correcting this misdiagnosis is crucial for understanding the adequacy conditions of causal models more generally.

Causal representations are distinguished from non-causal ones by their ability to predict the results of interventions. Interventions are typically modeled using the *do-operator*, which, when applied to a variable, breaks all of (and only) the causal arrows going into it. While the do-operator is a formal operation, it can be used to predict the outcome of physical interventions meeting certain causal conditions about how the intervention influences the intervened-upon variable relative to others. I will say that a model makes a *false prediction* about the results of interventions in cases where physical interventions on a variable in the model result in a new set of causal relationships that could not have resulted from applying the do-operator to that variable in the initial model. The *Manipulation Postulate* (Dash and Druzdzel 2001) is the requirement that causal models do not contain variables regarding which they make such false predictions. This postulate provides an adequacy condition on causal models, in the sense that models must obey the postulate in order to accurately describe the causal relationships in the system. Given the close conceptual connection between causation and interventions, as well as the role that the do-operator plays in fixing the causal content of a model, it is unclear in what sense a model that does not satisfy the postulate can still count as causal.

The manipulation postulate, combined with Dash's (2003) results, entails that for a specifiable class of dynamical systems, the causal models that would be inferred from sampling the variables at equilibrium will be inadequate. Contra Dash, I will argue that dynamic causal models reveal why one should reject the postulate. At issue in this dispute is the relationship between the do-operator and a distinct *equilibration* operator designed by Iwasaki and Simon (1994) to derive equilibrium models from dynamic ones. Applying the equilibration operator to a variable that is away from equilibrium yields a model of the causal relations that would obtain in the system were that variable to reach equilibrium. Certain dynamic causal models contain variables such that equilibrating them results in models violating the postulate, and applying the do-operator to them results in models that do not (Fig. 1). This would seem to be an excellent reason for rejecting the models derived via equilibration, as Dash does. Moreover, since these models are those that would be inferred by applying standard causal search methods to the variables at equilibrium (Dash 2003, pp. 60–66), they pose a problem not just for modelers applying the equilibration operator to a dynamic model, but for anyone modeling a system at equilibrium. In fact, this issue arises not just for equilibrium systems, but for any system sampled at a rate such that the variables have had sufficient time to reach a steady-state in response to perturbations.

The manipulation postulate misses that intervening is only half the picture. Just as it is possible to treat the do-operator as formalizing a physical intervention, the

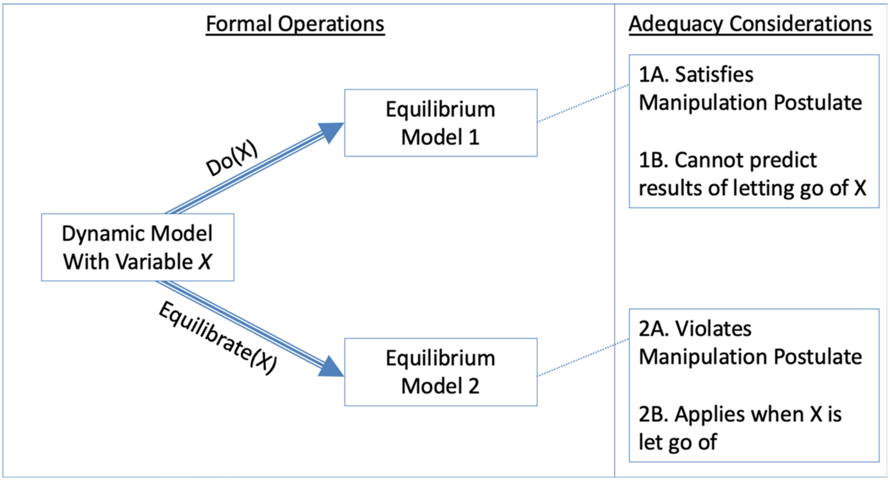

Formal Operations
Adequacy Considerations
Dynamic Model With Variable *X*
Do(X)
Equilibrate(X)
Equilibrium Model 1
Equilibrium Model 2
1A. Satisfies Manipulation Postulate
1B. Cannot predict results of letting go of X
2A. Violates Manipulation Postulate
2B. Applies when X is let go of

Fig. 1 High-level map of models, operations and adequacy considerations. Adequacy considerations concern the ability of the models to predict how the system will change in response to the actions modeled by the formal operations. Considering intervening without "letting go" leads one to illegitimately privilege 1A and 2A over 1B and 2B

equilibration operator can similarly be understood as modeling a physical action, which I call "letting go". Unlike the do-operator, the equilibration operator only results in non-trivial transformations when applied to dynamic causal models. But certain equilibrium models nevertheless describe the causal relationships resulting from letting go of a variable. In the class of cases being considered, the equilibrium models will either make false predictions about the results of intervening, or lack the resources to determine how the system will change as a result of letting go. I argue against the manipulation postulate on the basis that it makes an arbitrary distinction between these limitations.

The discussion here is compatible with explicating causal relationships in terms of interventions. Where the postulate goes wrong is in requiring that a model makes no false predictions regarding *any* of the variables in a model, including those that one need not intervene upon to test causal relationships. Certain causal relationships in a model obtain only when one does *not* intervene on particular variables. Dynamic causal models reveal this to be a general phenomenon—some systems will have one set of equilibrium causal relationships for the scenario where one intervenes on a particular variable, and a separate set of relationships for when one lets go of it (which requires not intervening upon it).

While this paper is structured as an argument for rejecting the manipulation postulate, much more interesting is the reason *why* it should be rejected. In talking about model adequacy, one can distinguish between (A) a characterization of the systems to which the models apply, and (B) what the models predict about the systems to which they apply.[1] For causation, there has been limited discussion of (A), beyond general claims that a system's causal relationships are relative to a "causal setup" (Hausman 1998, p. 25) or a "causal field" (Mackie 1974). The answer to (B) is more straightforward: models predict the results of interventions. What the manipulation postulate—and the broader literature—misses is that this is not the only role that interventions play. Physical interventions matter not just for establishing causal relationships in systems to which a model applies, but can also determine whether the model in fact applies to a system. As a result of being insensitive to this dual role of interventions, the manipulation postulate goes beyond requiring models to predict the results of interventions for the systems to which they apply, but rather arbitrarily legislates that only the models for certain systems are adequate.

2 Statics, Dynamics, and Causal Models

In this section, I explain how causal models describe the behavior of systems both at and away from equilibrium. I begin with with Simon's (1953) *causal ordering method* and explain how it represents a system's equilibrium behavior. I then consider *dynamic causal models*, which generalize this method to systems away from equilibrium. I introduce these models with a single example that can be modeled both statically and dynamically. Readers interested in further details may consult Simon and Rescher (1966), Iwasaki and Simon (1994), Dash (2003), Weinberger (2019, 2020), and the "Appendix" below.

Simon (1953) considered what makes causal relationships asymmetric, given that the equations stating scientific laws are typically symmetric. Consider the ideal gas law, which states that a gas' pressure times its volume is proportional to its temperature:

$$PV = kT \tag{1}$$

This equation is silent regarding the causal relationships among the variables. It does not indicate, for example, whether temperature causes pressure or pressure causes temperature (or neither). Yet, Simon says, the causal relationships are derivable given a *set* of equations. Suppose the values of *temperature* and *volume* in a system can be fixed independently of the other variables. Concretely, the gas might be in a sealed container immersed in a heat bath with constant temperature. Since these variables' values are independently set to constants, they can be represented with equations as follows:

$$T = c_2 \tag{2}$$

$$V = c_3 \tag{3}$$

[1] I am grateful to Michal Hladky for helpful discussion regarding this distinction.

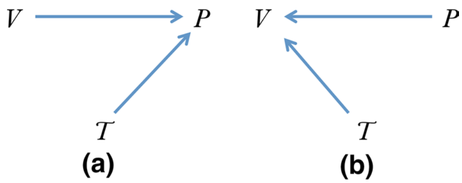


Fig. 2 Graphs for **a** fixed-volume system, **b** moveable piston system

In Eqs. (1)–(3), the values of T and V are given by (2) and (3) and given these values equation (1) determines P's value. Simon's account, in short, is that a set of symmetric equations can be written as a set of asymmetric equations when certain sets variables can be solved for before other sets of variables, in which case the former are causes of the latter. In this case, T and V are causes of P (Fig. 2a), and the ideal gas law in (1) can be rewritten as the asymmetric equation:

$$P = kT/V \tag{1a}$$

A *causal ordering* over the variables is a partial ordering in which effects come later than their causes.

Variables whose values are determined independently of the others—such as T and V in Eqs. (2) and (3)—are *exogenous*. Clearly, the causal ordering depends on which variables count as exogenous. Given a gas in a moveable piston, *volume* would not be exogenous. Rather, *pressure* would be:

$$P = c_4 \tag{4}$$

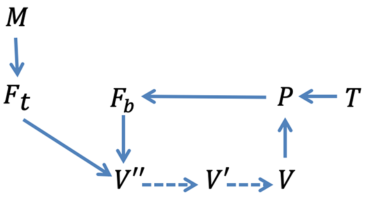
For Eqs. (1), (2), and (4), the causal relations are those in Fig. 2b. We see that Simon's method does not derive causal knowledge without causal assumptions, but rather clarifies the assumptions that jointly imply a causal ordering. Notably, when it is possible to uniquely solve for a causal ordering, the symmetric equations can be rewritten as *structural equations* in which each variable is given on the left-hand side of an equation in which its causes are on the right. While (1a) (along with (2) and (3)) is one of the structural equations for the fixed-volume system, the movable piston system would have the following structural equation:

$$V = kT/P \tag{1b}$$

While structural equations are often introduced with the stipulation that the variables on the left asymmetrically depend on those on the right, the causal ordering method reveals how to derive this asymmetry from a set of symmetric equations.

Simon's methods were a key step in the development of more recent (and better known) causal modeling methods (Spirtes et al. 2000; Pearl 2009). Simon's insight was that the existence of a causal ordering depends upon there being a set of equations in which distinct equations correspond to autonomous mechanisms.[2]

[2] The significance of autonomous mechanisms for causal ordering is defended in detail in Hausman (1998) and Hausman and Woodward (1999). Peters et al. (2017, p.16) provides a clear-cut example of the continued centrality of mechanism independence assumptions for causal inference.



(5) $M = m$

(6) $F_t = Mg$

(7) $V'' = (F_t + F_b)/m_p$

(8) $P = kT/V$

(9) $T = t$

(10) $F_b = rP$

(11) $V'_{t+1} = V'_t + V''_t \Delta t$

(12) $V_{t+1} = V_t + V'_t \Delta t$

Fig. 3 Dynamic causal model for the ideal gas system

The structural equations derived from his method are central both to causal inference from probabilities and interventionist theories of causation. Adding independent error terms to deterministic structural equations yields a probability distribution satisfying the *causal Markov condition*, a core assumption of graphical causal inference techniques. Additionally, and relatedly, structural equations are interpretable as indicating how effects will change given interventions on their causes.

In Sect. 6, I will argue that the models in Fig. 2 correctly describe the way that their effect variables respond to interventions on their causes. Perhaps controversially, I maintain that one can intervene on P in the moveable piston system by adjusting the mass on top of the piston. As my defense of this claim requires delving into thorny issues about interventions on equilibrium variables, I postpone it until after my main argument.

Iwasaki and Simon (1994, p. 145) claim that Simon's initial method was intended for "static system[s] of equilibrium equations". Such equations contain variables that are represented as being simultaneous, and there are several ways to interpret such simultaneous relationships (Malinsky and Spirtes 2018). In Iwasaki and Simon's framework, these equations represent the variables at a point at which they have reached steady-state. The use of simultaneous equations to represent causal relationships need not indicate simultaneous causation. Rather, such equations indicate that the variables adjust their values to one another so quickly that we can model the relationships as if they were instantaneous. See Anderson (2020) and Weinberger (2019, 2020) for further discussion.

The ideal gas law, and the causal models for systems instantiating it, refers to variables at *equilibrium.* In the moveable piston system, changing the temperature of the heat bath will change the gas' equilibrium volume, but not its pressure. In contrast, when the system has not yet reached equilibrium and volume is expanding, there will be a feedback loop by which *pressure* and *volume* influence one another.

The *dynamic causal model* for the system in which volume is away from equilibrium is given in Fig. 3 (see "Appendix" for derivation). As before, we have the variables *temperature* (T), *pressure* (P) and *volume* (V). Notably, the first- and second-time-derivatives of *volume* indicating its velocity, V', and acceleration, V'', are included in the model to indicate that V is not at steady-state, but rather changing over time.[3] There are also forces on the bottom (F_b) and top (F_t) of the piston, with F_t determined by the mass (M) on top of the piston and gravity (g). These forces combined determine the piston's acceleration, which in turn determines its velocity and then volume. The determination relations among V, V', and V'' are not causal relations, since they are mathematically related. These mathematical relationships incorporate the passage of time into the model. Through integration (taking the integral), one can use a variable and its derivative function to predict the variable's value at a subsequent time step. For instance, in Eq. (12) one predicts the value of V at a subsequent time step by integrating its derivative V' to determine how V will change over an interval Δt and combining this result with its prior value (a constant not supplied by integration).[4] Similarly, one can use the acceleration and velocity of V at a time to predict its velocity at a subsequent time. In the graph, the dashed arrows from higher- to lower-order derivatives are called *integration links*. While the graph with integration links has cycles, time-indexing the variables yields an acyclic graph (see "Appendix").

Introducing derivatives into a causal model changes it in subtle ways. Integration links are the most obvious innovation. While variables not linked by integration links are represented as influencing one another instantaneously, variables so linked influence one another at a slower rate (since the cause does not influence its effect within a time-step, but only after an arbitrarily small lag). So dynamic causal models distinguish among causal relationships occurring at faster and slower rates.[5] Less obviously, but just as important, when introducing a time-derivative for a variable one also needs to specify that variable's initial value. In equation (12), the value of volume at a time-step depends on its value at the previous time step, and its value at some initial time step must be specified in order to predict how it will evolve. The need for initial conditions marks an important difference between dynamic and equilibrium models. Equilibrium systems are "memory-less" in the sense that since variables have fully adjusted to any changes in their causes, the value of a variable is fully determined by its causes at that time step. In contrast, *volume* at a time-step depends on its value at the prior time step.

The crucial feature of the dynamic model for what follows is that one can derive either of the equilibrium models given in Fig. 2 from it . There are two formal operations one can apply to a dynamic model: *intervention* and *equilibration*. Interventions (also known as *manipulations*) set a variable to a constant value in a particular way (Sect. 4), and can be represented by the do-operator, which breaks the arrows

[3] Dash (2003) replaces *volume* with the variable *height* since volume is proportional to height for a fixed cross-section and height varies along only one spatial dimension. Here I keep the variable *volume* to avoid labeling the same variable differently across models.

[4] While the equations here use discrete time-steps, integration paradigmatically applies to continuous functions.

[5] When some variable X influences Y at a slower rate, and Y influences Z instantaneously, the influence of X on Z is also delayed. More generally, Y serves as a bottleneck (rate-limiting factor) delaying all downstream variables from adjusting to changes in X (thanks to Shannon Nolen for pointing this out). So dynamic causal models do not just distinguish between faster- and slower-occurring interactions, but also partition sets of variables such that variables within subsets influence one another locally and rapidly, while causal influences across subsets occur more slowly (Weinberger 2020).

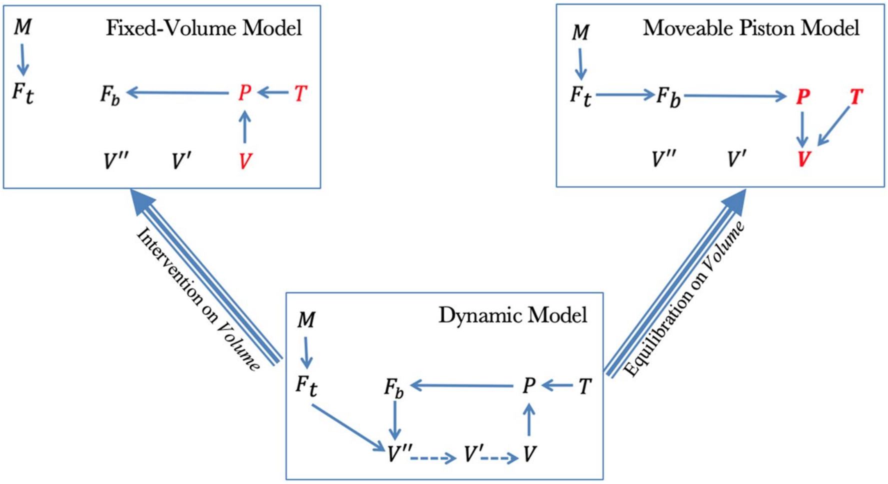


Fig. 4 Model map

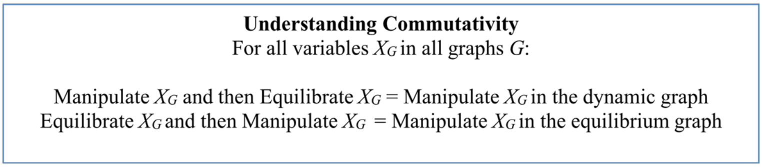
going into the intervened-upon variable. Interventions on a variable in a dynamic model further set its higher-order derivatives to zero. This makes sense—one cannot hold an object in place without giving it a velocity and acceleration of zero. Intervening on *volume* in the dynamic graph (Fig. 4) yields the same causal relations among P, V, and T, as those in the equilibrium model for the fixed-volume system (Fig. 2a).

Applying the equilibration operator to a variable whose derivative is in the model yields a model in which that variable has reached equilibrium.[6]Dash (2003) provides a schema for deriving the model resulting from equilibrating X (further details in “Appendix”):

1. Set all derivatives of X in the model to 0 and remove them from the model
2. Delete all equations going into X or its derivatives
3. Remap to get the new causal ordering

The second step involves deleting integration equations in addition to structural equations. Applying equilibration to *volume* in the dynamic model (Fig. 4) produces the same causal relationships among P, V, and T as those in the moveable piston system from Fig. 2b.

[6] In equilibrating multiple variables, one must equilibrate slower-equilibrating variables no later than faster-equilibrating variables (Iwasaki and Simon 1994, p. 166).



Understanding Commutativity
For all variables X_G in all graphs G:

Manipulate X_G and then Equilibrate X_G = Manipulate X_G in the dynamic graph
Equilibrate X_G and then Manipulate X_G = Manipulate X_G in the equilibrium graph

Fig. 5 Commutativity

3 The Red Herring of Commutativity

It is not obviously problematic that there are two distinct equilibrium models for the ideal gas system. In fact, existing discussions of such systems making no reference to dynamics propose these two models (Woodward 2003; Hausman et al. 2014; Woodward 2020). The different causal relationships reflect those obtaining when different factors are held fixed. Were one to add independent error terms the variables in each model, the causal models would satisfy the causal Markov condition for the resulting probability distributions, though they would be different distributions. As the different models correspond to distinct scenarios, there is nothing paradoxical about this.

So what's the problem? In Dash's dissertation (2003) and subsequent work (Voortman et al. 2012) he shows that in cases like our example, the equilibration and manipulation (i.e. "do-") operators do not commute. That is, one gets a different graph depending on the order in which one applies the operators. In focusing on commutativity, Dash is, of course, assuming that the operators *should* commute. This assumption is warranted for formal operations representing actions such that the order in which they are performed makes no difference to the net result. In this section I will explain why this is not a good assumption for the class of cases being considered.

Let's begin by getting a feel for the results of sequentially applying the two operators (Fig. 5). First, when one applies the do-operator and then the equilibration operator to a dynamic graph, the equilibration operator has no further effect on the graph. So manipulating and equilibrating yields the same result as just manipulating. Second, when equilibrating and then manipulating, it is not necessary to imagine that one starts with the dynamic graph and then derives the equilibrium graph from it via equilibration, as the latter is what one would infer when sampling the system at equilibrium. Accordingly, this order of operations amounts to applying the do-operator to the equilibrium graph. We see that the class of cases in which the operators do not commute is simply those in which a system's dynamic and equilibrium models differ in their causal orderings.

Differences in the causal ordering between dynamic and equilibrium models may be surprising, but they are not obviously problematic. Recall that the dynamic causal models can be "unfolded" into models with time-indexed variables ("Appendix"). While the causal relationships between the variables in the dynamic and equilibrium models appear to reverse, the variable sets being considered are not the same. To illustrate with a simpler example, turning on an oven in a room regulated by a thermostat may raise the room's temperature five minutes later, but not an hour later. While a shorter-scale model, but not a longer-scale model, might depict the oven as causing temperature, any apparent contradiction dissolves once one sees that the variables for temperature at different times are distinct.[7] In the moveable piston example, the differences between the dynamic and equilibrium models are more drastic than simply the presence or absence of a causal arrow. Nevertheless, the example establishes that differences in the causal relationships across the models do not imply that one must be wrong, as the models have different variables.

In the cases Dash considers, the explanation for the differences between the equilibrium and dynamic causal orderings is as follows. For any dynamic model containing a dynamic variable that influences its higher-order derivatives via some other variables, equilibrating that variable alters the causal ordering (Iwasaki and Simon 1994, p. 167; Dash 2003, § 2.3.1). Such variables are called *indirectly self-regulating* variables. In Fig. 3, for example, V influences its highest-order derivative V'' via P and F_b. Equilibrating V corresponds to allowing V to reach its equilibrium value via a process of self-regulation. The feedback loop in the dynamic graph is necessary in order for the system to reach equilibrium in this manner.[8] Intervening on V breaks this feedback loop and thus eliminates the conditions that were necessary for the system to equilibrate in this way, resulting instead in a distinct equilibrium state.

Once one sees what is going on, it becomes clear why the operators *shouldn't* necessarily commute. In his lone comment on why commutativity matters, Dash describes interventions on the dynamical graph as “shocks" (2003, p. 4) that push the system out of equilibrium. But interventions on dynamic graphs as conceptualized by the do-operator are not mere shocks that temporarily move the system away from equilibrium, but “clamp” interventions that fix the variable indefinitely and determine a *new* equilibrium.[9] That is, intervening on variables in the dynamic model is a *way* of bringing the system to an equilibrium state, and it is a different equilibrium state from that which would result from letting the variable reach a stable value in the absence of such an intervention. Since these are distinct equilibrium states of the system, commutativity is not warranted.

To the extent that Dash's results provide reason to worry about the adequacy of models derived from equilibration, it is not because they diverge in their ordering from the dynamic model. A greater cause for concern is the fact that the models derived via equilibration omit information that is relevant to predicting how the

[7] Such examples may provide counterexamples to the causal Faithfulness condition, which implies that variables linked by causal chains will be probabilistically dependent. What to say about this remains unresolved, as the literature on Faithfulness has focused on cancelling paths rather than damping. In any event, faithfulness should be understood as a helpful simplifying assumption for causal inference rather than a candidate adequacy principle for causal models.

[8] It is straightforward to show that in a representation of a system containing both X and X', X will reach a stable steady-state only if X' is a function of X (Dash 2003, pp. 37–38). In dynamic causal models, this function corresponds to the existence of a causal path from X to X' and such paths are disrupted by interventions on the feedback loop.

[9] Thanks to Jim Woodward and Chris Hitchcock for independently highlighting the importance of distinguishing between shock and clamp interventions in this context. To be clear, there is no ambiguity that Dash (2003, p. 28) models interventions on dynamic systems as “clamp” interventions.

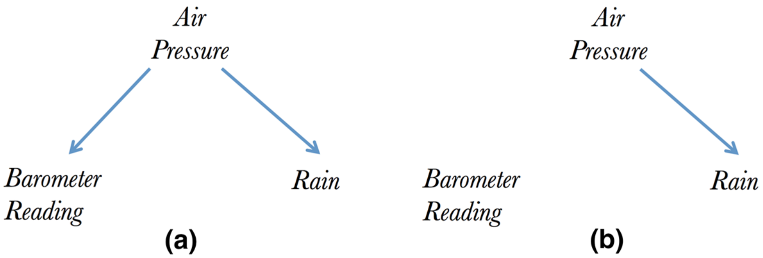


Fig. 6 **a** Causal graph for Barometer example, **b** an intervention on *Barometer Reading*

modeled system will change as a result of interventions. In particular, they do not represent the dynamic feedback loop that maintains the system at equilibrium, and thus cannot flag that this feedback loop is destroyed by interventions on variables in the equilibrium model. Moreover, the causal relations resulting from these interventions differ from those that could be derived by applying the do-operator to the equilibrium model. Consequently, in cases where the equilibrium ordering diverges from the dynamic one, the equilibrium model might be accused of failing in the most basic function of causal models—predicting the results of interventions. In the next section, I will briefly review the concept of an intervention prior to using Dash and Druzdzel's *manipulation postulate* to make this worry more precise.

4 Interventions and the Manipulation Postulate

A primary use of causal knowledge is for predicting the results of hypothetical interventions. Learning that users of a drug are less likely to get heart disease is useful only if intervening to give patients the drug would reduce their chance of heart disease, at least for some. Interventions are central to Woodward's (2003) account of causal explanation. On his account, *C* causes *E* if it is possible to change *E* via an *ideal intervention* on *C*. An intervention on *C* that is ideal with respect to *E* determines *C*'s value in a way such that any influence of *C* on *E* is via *C*. To give a standard example, a barometer does not cause rain, since although the barometer reading is correlated with future rain, one cannot change the chance of rain by ideally intervening on the barometer—say by installing a widget that holds its display at a low value. An intervention on the barometer would *fail* to be ideal if, e.g., one changed the barometer reading by (somehow) intervening on the atmospheric pressure. This would change both the barometer reading and the chance of rain, but would not influence the chance of rain *via* changing the barometer reading.

Woodward's account is in part motivated by graphical causal modeling methods (Spirtes et al. 2000; Pearl 2009). The causal graph for the barometer example is given in Fig. 6a. In it, the nodes are random variables and the directed edges (i.e. arrows) represent direct causal relationships. Direct causal relationships are often explicated in terms of whether one can change the effect variable via ideally intervening on a cause while holding all other variables in the model fixed. Ideal interventions disable the relationship between a variable and its prior causes, and thus can be represented as breaking the arrows into the intervened-upon variable, as in Fig. 6b. Alternatively, one can represent the intervention within the model with an intervention (or "policy") variable, which is a cause of the intervened-upon variable such that, for particular values of the intervention variable, the intervened-upon variable does not depend on its other causes in the model.

Given this background, Dash and Druzdzel's (2001, p. 194) *manipulation postulate* might seem almost trivial. It states that when a variable in a graph is manipulated, this manipulation will *at most* break the arrows going into the manipulated variable.[10] Manipulations that are properly represented by the do-operator automatically satisfy the postulate, as they break all of the arrows going into the intervened-upon variable. The postulate reflects the assumption that interventions on a variable only locally influence that variable, and that all influences on other variables are only *via* influencing that variable.

At first glance, the manipulation postulate appears to be simply clarifying the formal operation of an intervention, allowing for the fact that an intervention might be "soft" and thus not break the arrows into the intervened upon variable (Korb et al. 2004). But the manipulation postulate cannot be an analytic truth, since Dash and Druzdzel claim that it can be empirically violated. This presupposes that one can talk about physical interventions independent of whether they can be formalized using the do-operator. Schematically, to argue that the postulate is violated by G, one must (1) claim that a particular action that changes one (or more) of the variables in the system characterized by G should count as an intervention on that variable, (2) posit a graph G' that represents the causal relationships in the system resulting from this action, and (3) show that the edges in G' are not a subset of those in G. In cases where the postulate is violated, I will say that model G makes *false predictions* about the results on interventions or the relevant variables.

Arguing that a particular action on a variable ought to be characterized as an intervention on that variable is a subtle matter. Especially in cases where the action results in consequences that are incompatible with the application of the do-operator, there will be a temptation to argue that the action was not the sort of action that the do-operator was designed to model in the first place. But in order for causal models to make predictions about actions in the world, it must be possible to specify the conditions that an action must meet to be represented using the do-operator. Although there has been little discussion about how to determine whether a physical action ought to be represented as an intervention, these causal conditions are well-understood.[11]

Consider the requirement that interventions on a variable need to influence that variable's effects only *via* their influence on that variable. In reality, many actions will not meet this condition, but will be "fat handed", and such actions should not be

[10] Dash and Druzdzel's more precise version (slightly modified for exposition) is as follows: If graph G consisting of vertices (nodes) V and edges E is a causal graph, and $V' \subset V$ is a subset of variables being manipulated, then the causal graph G' for the manipulated system is such that $G' = \langle V, E \rangle$, where $E' \subseteq E$ and E' differs from E by at most the set directed edges into V'.

[11] See Prescott-Couch (2017) for a discussion that is sensitive to these issues.

represented using the do-operator. But whether an intervention is fat handed is not just a feature of the model, but depends on whether the physical intervention in fact influences downstream variables via an avenue not going via the intervened upon variable. Similarly, the requirement that the intervention be uncorrelated with causes of variables other than the intervened upon variable is also a claim about reality. Here I am not suggesting that it is trivial to determine when an action counts as an intervention, but only demonstrating that this is a meaningful question.

Note that the false predictions of models violating the postulate are not predictions about an effect's response to interventions on its causes. Rather they concern how the qualitative causal structure of the graph will change in response to interventions on variables. Failures of such predictions are especially worrisome, since interventions play a key role in spelling out the causal content of causal models. Spirtes, Glymour and Scheines' (2000, §3.7.2) manipulation *theorem* provides the conditions that license inferring the results of a hypothetical arrow-breaking intervention based on the passively observed (i.e., unmanipulated) probability distribution. This theorem is important for understanding the empirical content of causal models in cases where one does not experimentally intervene. Spirtes et al. (2000, p. 51) acknowledge that the manipulation theorem does not apply in cases where the direction of the causal relationship between two variables reverses. As the violations of the manipulation postulate described do involve reversibility, they fall under this stated exception to the manipulation theorem. But to the extent that the manipulation theorem spells out the empirical content of causal models, we need an account of the causal relationships in these exceptional cases.

Note that the issue with violations of postulate is not merely that many real world interventions are not aptly represented by the do-operator. Stern (2019) provides cases involving uncertainty regarding whether physical actions should be represented as "arrow breaking" interventions represented by the "do-operator" as opposed to soft interventions. Since even soft interventions preserve causal structure, such cases do not violate the postulate.

Violations of the postulate differ from certain less problematic failures of models to predict interventions. It is widely acknowledged that causal models posit relationships that obtain only under a range of interventions on the causes. For instance, Hooke's law, which describes the force exerted by a stretched spring, can be interpreted causally even though it no longer applies when the spring is stretched too far. Additionally, causal relationships only obtain given certain background conditions—striking a match causes it to ignite only in the presence of oxygen. These types of cases are much less problematic than those involving violations of the postulate. While they suggest that even adequate causal models will apply only in a limited range of contexts, there is no issue of the causal interpretability of the relationships in the situations to which the models apply.

Dash and Druzdzel (2001) assert that "[a]ll formalisms for causal reasoning take the Manipulation Postulate as a fundamental starting point" (194). While they are the only writers I know of to label the postulate, it is exceedingly plausible and is widely assumed. In causal models, the omitted arrows are as important—arguably more so—as those that are included. If there is no arrow from X to Y, it should not be possible to change Y via intervening on X. In violations of the manipulation postulate involving interventions on X, such interventions directly influence variables that are not direct effects of X in the model.[12] For lack of a better term, I refer to the postulate as an "adequacy criterion". By this I mean a proposal for how a model must relate to its target system in order to count as accurately describing the system's causal relationships. While there clearly is a close connection between a model's being adequate and its predicting the results of interventions, I will argue that the link provided by the manipulation postulate is untenable.

5 Why the Manipulation Postulate Should Be Rejected

To see why the equilibrium graph for the moveable piston (Fig. 2b) violates the manipulation postulate, imagine that we intervene on the piston to fix its volume. We might do this by inserting a pin into the side of the gas container to hold the piston in place (cf. Hausman et al. 2014). This counts as a physical intervention on volume, because it holds volume fixed, and, for a given temperature, the equilibrium pressure of the system depends on the particular volume to which one holds it fixed. We see that any effect on the other variables is only indirectly via volume. If Fig. 2b were the correct graph, then modeling this intervention using the do-operator yields the prediction that this intervention will at most break the arrows going into V. But this is not the case. As a result of the intervention, the causal relations are those given in Fig. 2a. As expected, the arrows going into V are no longer there. But now there are also arrows from V and T to P that were not in the initial model. We see that the equilibrium graph for the movable piston violates the manipulation postulate with respect to interventions on V.

The graph for the sealed container makes no similar false prediction, and thus does not violate the postulate. It captures how volume and temperature can be exogenously fixed, and the values to which they are fixed determines equilibrium pressure. More generally, in all cases where equilibrating and manipulating the dynamic graph yield different models, only the models derived from equilibration violate the postulate.

Despite its plausibility, the manipulation postulate serves as an arbitrary basis for differentiating the equilibrium models. Imagine that instead of inserting the pin into the moveable piston to hold the volume fixed, one removes the pin that fixes the volume and allows the volume to vary. Then, I claim, the correct graph for the system is not that in Fig. 2a, but rather that in Fig. 2b. The intuitive defense of this claim is that were one to determine equilibrium pressure by adjusting the weight on top

[12] To illustrate, suppose that $X \rightarrow W \rightarrow Z$ and the postulate fails due to an intervention on X that breaks the arrow $W \rightarrow Z$. To influence this relationship, X must influence Z through a path not going through W. So X's altering the relationship between W and Z amounts to its directly influencing Z, despite the absence of a direct connection in the model.

of the piston, this would count as an intervention on pressure, and the equilibrium volume of the system would then depend on the values to which pressure and temperature are exogenously fixed. I'll offer a more detailed defense in the following section.

Just as the moveable piston model cannot predict the causal relationships that result from inserting the pin, the fixed-volume model cannot predict the causal relationships that result from removing the pin. Because inserting the pin can be represented using the do-operator and removing the pin cannot, only the former model makes a false prediction about the results of the action. But this does not reflect a difference in the adequacy of the models, but rather a blind spot in the formalism. One model cannot predict the results of intervening to hold volume fixed, and the other cannot predict the result of the symmetric action of "letting go" of volume and allowing it to vary. While, for equilibrium models, only one of these actions is captured by a formal operator—and thus only one model makes a *false* prediction—it remains the case that there are two mutually exclusive actions that one can perform on a variable, and each model captures the results of only one of them.

In dynamic causal models, the action of letting go is captured by the formal operation of equilibration, which yields the causal relationships that would result were one to let the equilibrated variable reach steady-state without intervening upon it. Of course, whether applying equilibration yields the correct causal relationships is precisely what is at issue here. Nevertheless, once one acknowledges that equilibration—just as much as the do-operator—can be taken to represent a particular type of action performed on the system, it becomes clear that interventions cannot be treated as a neutral basis for distinguishing between the adequacy of the equilibrium models. Intervening and letting go correspond to two ways of bringing the system to two distinct equilibrium states that, at least potentially, involve two distinct sets of causal relations. Adopting the manipulation postulate amounts to arbitrarily privileging the model for one of these scenarios.

Contra Dash, dynamic causal models do not serve as a basis for choosing between equilibrium models. Rather, they reveal why there will be two distinct sets of models in a particular class of cases. The reason models derived via the equilibration operator fail to predict the results of intervening on the equilibrated variable is that doing so disrupts the feedback loop required for the system to naturally reach equilibrium. But the failure is symmetric. Note that the model for the fixed-volume system does not merely indicate what would happen *were* one to intervene, but rather only applies under the assumption that one *does* intervene to hold the volume fixed. This is not transparent from the equilibrium model itself, but would be apparent from seeing how the model is derived via applying the do-operator to the dynamic graph. This is precisely analogous to the way that one could use the dynamic causal model to infer that one *cannot* intervene on the self-regulating variable in the moveable piston model.

For the sake of rebutting the manipulation postulate, it is not necessary to argue that both equilibrium models are adequate. What matters is that they are on a par in terms of adequacy. One might be inclined to argue that in fact *neither* of the models are adequate since they either fail to predict the results of intervening or the results of letting go. This would be unappealing, because one of the equilibrium models is derived via applying the do-operator to the dynamic graph and saying that such interventions could yield inadequate graphs would throw even the adequacy of *dynamic* causal models into question. This is a stark illustration of why, whatever one says about the relative merits dynamic models over equilibrium ones, the difference between them is not properly understood by emphasizing the failure of certain equilibrium models to predict the results of interventions.

Rejecting the manipulation postulate is compatible with maintaining that adequate causal models must be able to predict the results of interventions. Although the model for the moveable piston cannot predict the results of interventions on volume, it does correctly depict how volume depends on temperature and pressure. The problem with the postulate is its requiring that one can intervene upon *any* variable in a causal model. The ability of certain equilibrium models to predict the interventionist-causal relationships that they do depends on one's *not* intervening on certain variables. At first, this defense of the adequacy of equilibrium models might seem ad hoc, especially since the model does not specify which variables one should not intervene upon. Yet the present discussion reveals why requiring causal models to predict the results of all interventions is too high a bar, and leads one to make arbitrary distinctions between causal models.

6 Some Subtleties Regarding Interventions

The discussion so far has presupposed that the equilibrium model for the fixed-volume system satisfies the manipulation postulate and that the equilibrium model for the moveable piston correctly predicts the results of interventions on the exogenous variables. But whether a particular physical action counts as an intervention can be a subtle matter. Here I will address worries about whether the representations of these interventions on the system is correct. Doing so will reveal that the prior discussion of the running example has glossed over some important subtleties regarding how to understand the interventions involved, but I will argue the success of my argument is invariant across distinct ways of addressing these subtleties.

There exist two traditions for modeling interventions. The dominant one models interventions as changing the values of variables, while there is an alternate tradition (e.g. Iwasaki and Simon 1994; Hoover 2001) that models interventions as changing parameters. In the latter treatment, for X to cause Y is for it to be possible to change the value of Y by intervening on the parameter in the structural equation giving Y as a function of X. This change is reflected in the fact that for a fixed value of X the value of Y will vary depending on the parameter value. These two approaches are sometimes treated as interchangeable,[13] although the difference between them

[13] Hoover (2013) provides the most thorough comparison of his own generalization of the Simon approach and the approaches of Woodward and Pearl, emphasizing the differences between accounts of causality based on the different approaches. See Malinsky (2018) for a recent assertion that "one may consistently adopt either framework for a given analysis" (2304). White and Chalak (2009) further develop the notion of an intervention on a parameter. Changing the parameter linking X to Y can be modeled as a soft intervention on Y.

potentially matters for explicating the equilibrium relationships between P on V in the different models, as I now explain.

It is essential for my argument that although the graph for the moveable piston violates the manipulation postulate, it nevertheless predicts the response of V to interventions on P and T. Earlier I suggested one could intervene on pressure by placing a mass on top of the piston. This does, in fact, determine what the equilibrium pressure will be, which in turn determines the equilibrium volume (along with temperature). Nevertheless, considering the physical interpretation of the relevant quantities might shake one's confidence in this description. The pressure of the gas corresponds to the force exerted on the bottom of piston per unit area. Placing a mass on top of the piston increases the force on the top of the piston. Since the factors causing these forces are distinct, it seems strange to say that the force on the top is a cause of the force on the bottom (and thus pressure). Dash and Druzdzel (2001, p. 196) describe this relationship as follows: since at equilibrium the force on the bottom of the piston must equal the force on top, fixing the force on the top determines what the force on the bottom will be at equilibrium. This explanation would suffice to establish this relationship as causal for someone who accepts the causal ordering method, but the equilibrium models produced by this method are at issue here.

The most direct way to address these concerns about the moveable piston model would be to show that placing the mass on top of the piston influences volume only indirectly via pressure. But this runs into complications. If we consider the volume and pressure of the system *away* from equilibrium, then the dynamic graph reveals that the mass does in fact influence pressure via volume. This by itself does not settle the question of whether the mass influences *equilibrium* pressure via influencing *equilibrium* volume.[14] But what *does* settle it? As we are only considering the equilibrium values of these variables, how do we determine whether certain interventions influence one only via the other? Note that this concern also applies to the fixed-volume system, since one needs to clarify why inserting the pin influences P via V, given that it determines the equilibrium values of both.

We find ourselves in the uncomfortable situation of lacking an example of a physical action that indisputably should count as an intervention on pressure. How, then, might we convince those who do not already accept the proposed equilibrium model for the moveable piston system? Here is where the notion of an intervention on a parameter comes into play. By showing that one can change the value of V by altering the parameter by which P influences V we can provide an independent method for justifying the model.

What does it mean to alter a causal parameter? According to Weinberger (2018), an intervention on the parameter a in the structural equation $Y = aX + U_Y$ in one model corresponds to an intervention on a variable Z in an augmented model in which Z is a cause of Y that influences the magnitude of Y's dependence on X. To

[14] One might think this question would be settled by assuming the transitivity of causation, though I take this assumption to be contentious in the context of dynamical systems that are subject to dampening forces.

illustrate, imagine one were measuring the effect of changing V on P across various fixed-volume systems immersed in heat baths of different temperatures. The magnitude of the effect of V on P will depend on the temperature of the heat bath. But if T is not included in the model, its influence will still be captured by a parameter in the structural equation by which V influences P.[15] This parameter, by design, depends on the value of T, so talk about variation in the parameter in the model with just P and V may be reinterpreted as variation in T in the augmented model containing it. Continuing along these lines, we see that testing whether V causes P by intervening on the parameter in the structural equation amounts to showing that one can change the influence of V on P by altering T.

But doesn't the claim that one can alter the parameter in the equation for P *presuppose* that V causes P? So what determines that varying T changes a parameter in the $V \rightarrow P$ relationship as opposed to a $P \rightarrow V$ relationship. The key fact is that in the fixed volume system V's value is determined *independently* of T.[16] This independence ensures that if altering T influences the relationship between P and V, it must be a cause of P (since it is not a cause of V). By analogous reasoning, we can establish that in the moveable piston setup, pressure causes volume, since pressure is set independently of temperature: placing the mass on top of the piston determines what the equilibrium pressure will be (as it will exactly counteract the force due to its weight), but even given this pressure, the equilibrium volume will still also depend on the temperature. It follows from this independence that by changing temperature one can intervene upon a parameter by which P causes V in a non-augmented model containing just P and V, and thus that $P \rightarrow V$.

We see that the causal relationships in the equilibrium graphs for the system can be explicated in terms of interventions on parameters. Yet the earlier discussion—including the formulation and criticism of the manipulation postulate—focused on interventions on variables. This is not necessarily a problem. The two notions of interventions can be seen as two distinct ways of characterizing the same causal structure.[17] Accordingly, the possibility of explicating the effect of V on P in terms of how it depends on the setting of T validates our prior judgment that fixing the volume of the system by inserting a pin into the container counts as an ideal intervention on V.

Yet this way of explicating the interventions on the moveable piston system might produce a novel problem for my claims about the fixed-volume model. I asserted that there is no formal operator corresponding to the action of removing the pin. Since the effect of P on V can be explicated as an intervention on a parameter, one might suggest that this action is in fact an intervention on P. Whether it is is unclear, since although one can intervene on pressure by adjusting the mass, it does not follow

[15] Depending on which systems are being represented, this parameter will reflect either the constant temperature across heat baths with the same temperature or an average temperature across systems with different values of T.

[16] The idea that the asymmetric dependency of variable Y on X depends upon the existence of independent causes of Y traces back to Hausman (1998).

[17] This is analogous to the way that Pearl's (2009) back-door and front-door criteria can each serve as bases for identifying a causal effects, even though the former works by conditioning upon causes of the cause and the latter requires one to condition on intermediate variables between the cause and effect.

that removing the pin also counts as an intervention. But the matter is sufficiently uncertain that it is worth considering this possibility. If removing the pin counts as an intervention on *P*, then the fixed-volume model violates the postulate, since the set of causal relations resulting from removing the pin are not those predicted by applying the do-operator. If so, then *both* equilibrium models violate the postulate.

Although this alternative framing would require changing some details of the argument, the net result preserves the conclusions of the previous section. First, as already noted, someone invoking the manipulation postulate as a basis for criticizing equilibrium models cannot reject this equilibrium model, as it is derived from the dynamic model via intervention. Accordingly, if the equilibrium model for the fixed volume system itself violates the postulate, then the postulate becomes self-undermining. Second, were one to attempt to model the removal of the pin as an intervention variable, this variable would serve as a common cause of pressure and volume. This is because the act of removing the pin excludes the act of inserting it (which would be another value of the variable) and inserting it counts as an intervention on volume. An action that counts as a common cause of two variables cannot be an ideal intervention on either of them. But beyond this technical point, the observation that the actions of intervening on and letting go of the pin are mutually exclusive and jointly exhaustive, yet cannot be represented using an ordinary intervention variable, reinforces my position that it is illegitimate to insist that only the model for the intervention scenario is adequate.

I have not provided a rigorous defense of the interchangeability of interventions on parameters and variables. In fact, one might take the need to switch to talking about interventions on parameters in order to independently motivate our equilibrium models as evidence that the two ways of talking are *not* interchangeable. This would open the door to further research into their relationship, and then a proponent of the intervention-on-variables approach might invoke the difficulty of unambiguously characterizing the interventions on the equilibrium models' exogenous variables as a novel basis for rejecting equilibrium models. For the purposes of this paper, however, two important points emerge from the this section's discussion. First, interventions on parameters provide a basis for motivating the causal relations in the equilibrium models. This is important, because without some basis for accepting the equilibrium models in the first place, the result that the manipulation postulate provides the *wrong* reason for rejecting them would be of little practical significance. Second, despite the existence of multiple ways of understanding the causal relations in the models and of characterizing the different interventions that influence them, there is no consistent characterization on which the manipulation postulate serves as a legitimate basis for privileging certain models as adequate.

7 Understanding Dynamic Causal Models

This paper has largely focused on what dynamic causal models do *not* do. They do not serve as a basis for distinguishing adequate from inadequate equilibrium models, where the former are those whose causal ordering matches that of the corresponding dynamic model. But the fact that they fail to serve this role does not diminish their significance. Dynamic models contain important information that is relevant to understanding a system's equilibrium behaviors. The information that an equilibrium graph was derived by equilibrating an indirectly self-regulating variable entails that the model will make false predictions about the results of interventions on that variable. My defence of the adequacy of such graphs does not mean that their inability to predict the results of certain interventions—or to provide any indication of this inability within the model—is not a practical limitation. Dynamic causal models address this limitation by enabling one to flag the variables that one cannot intervene upon in the corresponding equilibrium models while preserving those models' causal relationships. Symmetrically, they can be used to identify the variables that must be held fixed rather than let go of.

In the same way that standard causal models have been seen as tools for predicting the results of interventions, dynamic causal models should be understood as tools for predicting the results of interventions and equilibrations. Both of these operations correspond to ways of bringing the system to equilibrium. Accordingly, the adequacy conditions for *dynamic* causal models need to be understood in terms of their ability to predict certain equilibrium behaviors—specifically, those reflected in the equilibrium models. To be clear, in tying the adequacy of dynamic models to their ability to predict certain equilibrium behaviors I am not falsely claiming that dynamic models *only* predict the behaviors of systems at equilibrium. Rather, I am claiming that it is the structural features of the model—that is, qualitative features of the causal graph—that determine its adequacy. These features are operationalized in terms of how the model will change given applications of the intervention and equilibration operators. By analogy, to say that the adequacy of (standard) causal models depends on their ability to predict interventions is not to deny that they can explicate causal relations in both experimental and observational contexts.

Once one moves away from seeing dynamic models as arbitrating between adequate and inadequate equilibrium models, one can instead view them as playing a unifying role. In the class of cases considered, one can derive either of the equilibrium models by applying the do-operator or the equilibration operator to the relevant variable in the dynamic graph. Hausman et al. (2014) argue that there is no single graphical representation of the fixed- and variable-volume graphs for the ideal gas system. While they are correct that there is no single equilibrium graph, the dynamic graph for the system can be seen as providing precisely such a representation, as one can infer either set of equilibrium relationships from it. But this requires understanding the dynamic model in the way proposed here, rather than as a basis for choosing between the equilibrium models.

I have focused here on Iwasaki and Simon's framework, which remains one of the few systematic treatments of dynamic and equilibrium models. One thing that has emerged clearly from the discussion is that Iwasaki and Simon's dynamic models are closely tied to the behavior of a system at equilibrium. One might have expected a dynamic model to represent the broader dynamics of the system without making assumptions about its equilibrium behavior. But that is not what these dynamic models are doing. An advantage of Iwasaki and Simon's approach is that it is a generalization of Simon's earlier account, which remains an important framework for thinking about the structural equations in contemporary causal models. Additionally, it should be emphasized that reliance on assumptions about the longer-term steady-state behaviors of a system is ubiquitous in dynamical modeling. As Wilson (2017) argues, applying differential equations to model concrete systems typically requires a slew of additional assumptions, many of which appeal to the steady-state behavior of the system.

Given that Iwasaki and Simon's work is over twenty years old—an eon in academic time—the reader would be forgiven for questioning whether it is still relevant. There has, in fact, been a recent flurry of excellent work high-quality research on causation in equilibrium systems and on the relationship between standard causal representations and differential equations (see e.g. Mooij et al. (2013); Bongers and Mooij (2018)). A survey of the bibliography cited in this rapidly growing literature provides evidence that between Iwasaki and Simon (1994) and Mooij et al. (2013), work on this topic was relatively sparse.[18] Within this more recent literature, the work by Tineke Blom (Blom et al. 2018, 2020; Blom and Mooij 2021) is especially notable in taking Iwasaki and Simon's framework as as starting point.[19] Having an accurate picture of the relationship between dynamic and equilibrium models within this frameworks is thus crucial for evaluating recent developments in causally modeling dynamical systems.

As a final point, we must ask: to what extent does the discussion here generalize beyond the ideal gas case? Here, as elsewhere, Dash's dissertation helps. In the final chapter, he shows that the class of cases in which equilibration alters the causal ordering includes many paradigm physical systems, including simple harmonic oscillators, bodies in viscous media, and inverting amplifiers. The discussion here generalizes to those examples as well, since in each of them intervening and equilibrating produce different graphs. The ideal gas case is a useful starting point, since both equilibrium systems are familiar. By way of contrast, in the example involving a body submerged in a viscous medium, the equilibrium state in which the (upward) buoyant and (downward) gravitational forces cancel is more commonly discussed than the equilibrium state in which one intervenes on the body by holding it still. For this reason, the example is less useful for illustrating the difference between models derived from the two operations, but in some ways as effective at showing why we ought not take only equilibrium models derived by intervention as adequate.

[18] As further evidence, consider the references for research on causally modeling dynamic systems in Eberhardt (2017, p. 12). Three out of five sources refer to work by Dash and collaborators.

[19] Although here is not the place for a proper comparison between my treatment and Blom et al.'s, I will briefly flag one point of agreement and one point of disagreement. The discussion in Blom et al. (2018) agrees with the present discussion insofar as it assumes that there can be multiple distinct and adequate representations of systems at equilibrium, each of which applies given different "constraints". However, Blom et al. (2020); Blom and Mooij (2021) deny the causal interpretability of both equilibrium *and* dynamic causal models of the sorts considered here. This is of course incompatible with both the present discussion and the rest of the literature and demands further scrutiny.

8 Conclusion

In this article, I have argued that causal models can be adequate despite making systematically false predictions about the results of interventions on certain variables. The seemingly innocuous requirement that adequate models make no such predictions yields arbitrary distinctions between causal models. This is because certain equilibrium causal models only obtain when one either intervenes on a particular variable to hold it fixed, or lets go of it to allow variation, and the models do not internally cordon off these variables from others. Dynamic causal models do provide a basis for predicting the effects of intervening or letting go of such variables. But this does not undermine the adequacy of equilibrium models. All models apply only to particular systems given particular background assumptions, and causal models are no different. With the aid of dynamic causal models, I have offered insights into the nature of these assumptions and clarified their relation to the semantics of causal models.

Appendices

The equations for the dynamic model for the ideal gas system are as follows (Dash 2003)(numbers match those from Fig. 3 above):

$$M = m_0 \tag{5}$$

$$F_t = Mg \tag{6}$$

$$V'' = (F_t + F_b)/m_p \tag{7}$$

$$P = kT/V \tag{8}$$

$$T = t_0 \tag{9}$$

$$F_b = rP \tag{10}$$

$$V'_{t+1} = V'_t + V''_t \Delta t \tag{11}$$

$$V_{t+1} = V_t + V'_t \Delta t \tag{12}$$

Equations (5) and (9) specify that the mass (M) on top of the piston and the temperature of the heat bath are exogenous. Equation (6) says that the force exerted on top of the piston equals the mass of M times gravity g. Equation (7) uses Newton's second law to derive the acceleration of the piston by combining the forces on the top and bottom of the piston to get the net force, which is divide by the mass of the piston m_p. (8) is the ideal gas law. (10) states that F_b is proportional to P. (11) and (12) give

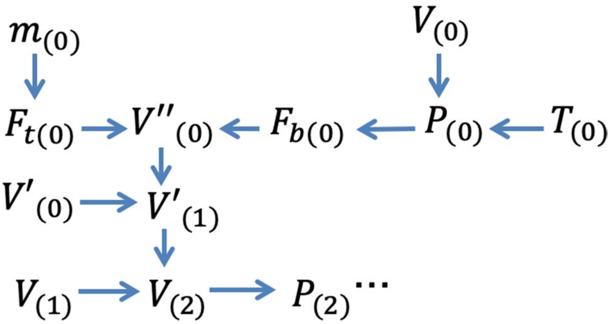


Fig. 7 Rolled out graph for the dynamic model in Fig. 3—subscripts in parentheses indicate time steps

V' and V at a time as a function of their values at the previous time-step and the next highest-order derivative at that time-step multiplied by the length of the time-step.

Because using (11) and (12) require the values of V and V' at a time in order to use their higher-order derivatives to predict their values at a subsequent time, the initial values of these variables, V_0 and V'_0, must be specified exogenously. This matters for the causal ordering and can be represented with the following equations:

$$V'_0 = v'_0 \tag{13}$$

$$V_0 = v_0 \tag{14}$$

We now derive the dynamic graph (Fig. 3). From V_0 (13) and T (9), Eq. (8) yields the value of P, which, from (10) yields F_b. M is exogenous (eq. (2)) and combined with (6) yields F_t. From (7) one can then derive V'' and then use equations (11) and (12) to derive V' and V at subsequent time steps. Note that while the equations were presented in the paper to reflect the causal ordering, the method just applied did not rely on information about which variables were on which sides of the equals signs.

As noted in the main text, although the dynamic graph appears to be cyclic, there exists a non-cyclic representation with time-indexed variables (Fig. 7). In Fig. 7, the ordinary causal relationships from the dynamic graph are represented synchronically, while variables connected by integration links are diachronically related.

Equilibration of V works as follows. In step 1, all of V's derivatives in the models are set to zero in the equations and removed as variables from the models. This means that (7) is replaced with:

$$0 = (F_t + F_b)/m_p \tag{7'}$$

(11) and (12) also contain derivatives of V, but this is moot, since in step 2 both these equations are deleted. More generally, while we have been describing Dash's informal sketch of equilibration, in his more rigorous characterization (2003, p. 30) he specifies that one must delete the equations for V and all of its higher-order derivative *except* its highest order derivative, which gets replaced with 0. In step 3, one

uses the remaining equations to re-solve for the causal ordering. Note that M is still exogenous (5) and still determines F_t from (6). From (7’) one can then derive F_b F_b combined with (10) yields the value of P. Since T remains exogenous (from (9)), one can then use the ideal gas law (8) to derive V from P and T.

Acknowledgements I developed my thoughts on dynamic causal models while participating in an online reading group with Jonathan Livengood, Karen Zwier, Shannon Nolen, and Adam Edwards. This paper has greatly benefitted from the comments of Matteo Colombo, David Danks, Nic Fillion, Paola Hernández, Edouard Machery, Shannon Nolen, Jim Woodward, and Karen Zwier on earlier drafts. I received helpful feedback from audiences in Düsseldorf, Etelsen, Tilburg, and Venice, and had useful discussions with Malcolm Forster, Gerhard Schurz and Mark Wilson. I am grateful to the Center for Philosophy of Science for funding the postdoctoral fellowship during which I worked out the core arguments in this paper, and to the Alexander von Humboldt Foundation for funding me during the period in which I revised it to its present form.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by a grant by the Humboldt Foundation

References

Anderson, W. (2020). The compatibility of differential equations and causal models reconsidered. *Erkenntnis, 85*, 317–332.

Blom, T., & Mooij, J. M. (2021). Causality and independence in perfectly adapted dynamical systems. https://arxiv.org/abs/2101.11885.

Blom, T., Bongers, S., & Mooij, J. M. (2018). Beyond structural causal models: Causal constraints models. In R. P. Adams, & V. Gogate (Eds.), *Proceedings of the 35th Annual Conference on Uncertainty in Artificial Intelligence (UAI-19)* (Vol. 15, pp. 440–448). Curran Associates, Inc.

Blom, T., van Diepen, M. M., & Mooij, J. M. (2020). Conditional independences and causal relations implied by sets of equations. arXiv preprint arXiv:2007.07183

Bongers, S., & Mooij, J. M. (2018). From random differential equations to structural causal models: The stochastic case. arXiv preprint arXiv:1803.08784

Dash, D. (2003). *Caveats for causal reasoning with equilibrium models*. University of Pittsburgh. PhD thesis.

Dash, D., & Druzdzel, M. (2001). *Caveats for causal reasoning with equilibrium models*. In S. Benferhat, & P. Besnard (Eds.), *Symbolic and quantitative approaches to reasoning with uncertainty. ECSQARU 2001. Lecture notes in computer science* (Vol. 2143). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-44652-4_18

Eberhardt, F. (2017). Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics, 3*(2), 81–91.

Hausman, D. M. (1998). *Causal asymmetries*. Cambridge University Press.

Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science, 50*(4), 521–583.

Hausman, D. M., Stern, R., & Weinberger, N. (2014). Systems without a graphical causal representation. *Synthese, 191*(8), 1925–1930.

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

Hoover, K. D. (2013). Identity, structure, and causal representation in scientific models. In H. K. Chao, S. T. Chen, & R. Millstein (Eds.), *Mechanism and causality in biology and economics* (pp. 35–57). Springer.

Iwasaki, Y., & Simon, H. A. (1994). Causality and model abstraction. *Artificial intelligence, 67*(1), 143–194.

Korb, K. B., Hope, L. R., Nicholson, A. E., & Axnick, K. (2004). Varieties of causal intervention. In *Pacific Rim international conference on artificial intelligence* (pp. 322–331). Springer.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford University Press.

Malinsky, D. (2018). Intervening on structure. *Synthese, 195*(5), 2295–2312.

Malinsky, D., & Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal discovery* (pp. 23–47).

Mooij, J. M., Janzing, D., & Schölkopf, B. (2013). From ordinary differential equations to structural causal models: the deterministic case. arXiv preprint arXiv:1304.7920.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.

Prescott-Couch, A. (2017). Explanation and manipulation 1. *Noûs, 51*(3), 484–520.

Simon, H. (1953). Causal ordering and identifiability. In W. Hood & T. Koopmans (Eds.), *Studies in econometric method* (pp. 49–74). Wiley.

Simon, H. A., & Rescher, N. (1966). Cause and counterfactual. *Philosophy of Science, 33*(4), 323–340.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT Press.

Stern, R. (2019). Decision and intervention. *Erkenntnis, 84*(4), 783–804.

Voortman, M., Dash, D., & Druzdzel, M. J. (2012). Learning why things change: the difference-based causality learner. arXiv preprint arXiv:1203.3525

Weinberger, N. (2018). Faithfulness, coordination and causal coincidences. *Erkenntnis, 83*(2), 113–133.

Weinberger, N. (2019). Reintroducing dynamics into static causal model. In S. Kleinberg (Ed.), *Time and causality across the sciences.* Cambridge University Press.

Weinberger, N. (2020). Near-decomposability and the time-scale relativity of causal representations. *Philosophy of Science, 87*, 841–856.

White, H., & Chalak, K. (2009). Settable systems: An extension of pearl's causal model with optimization, equilibrium, and learning. *Journal of Machine Learning Research, 10*(8), 1759–1799.

Wilson, M. (2017). *Physics Avoidance: And other essays in conceptual strategy*. Oxford University Press.

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

Woodward, J. (July 2020). Flagpoles anyone? Causal and explanatory asymmetries. http://philsci-archive.pitt.edu/17419/.