Homeostasis and the Faithless Foundations of Causal Inference

Naftali Weinberger

June 29, 2023

Abstract

The causal faithfulness condition, which licenses inferences from probabilistic to causal independence, is known to be violated in dynamical systems exhibiting homeostasis. Using the example of the Watt governor, I here present a precise causal characterization of such violations, which differ from cases involving cancelling paths. Traditional defenses of faithfulness do not carry over to cases of homeostasis and one should expect such cases to be widespread. Epistemically, this does not pose a problem for causal inference, since such failures of faithfulness do not undermine the ability of models at different time scales to properly represent the causal and probabilistic independence relations effectively obtaining at the relevant time scale. These failures are nevertheless significant, since they provide the key to understanding how causal attributions can apply within complex systems exhibiting high degrees of mutual dependence among their parts.

1 Introduction

Correlation does not imply causation. But does causation imply correlation? The causal faithfulness condition, which has played a central role in algorithms for choosing among graphical causal models, claims that the answer is "yes". But causation does *not* imply correlation. Paradigm counterexamples to the causal faithfulness condition ("faithfulness" for short) involve cancelling causal paths. If I exercise to burn calories, but the exercise leads me to eat more to make up the lost calories, then the exercise might have zero net impact on calories, even though exercise does burn calories. So faithfulness is not true in general.

Spirtes et al. (2001) defend faithfulness on the grounds that exact cancellations of causal paths are exceedingly improbable, although this argument remains controversial (Andersen, 2013; Steel, 2006). Zhang and Spirtes (2008) prove that faithfulness sometimes has testable implications, and develop algorithms that avoid outputting the wrong model in cases where faithfulness fails (Spirtes and Zhang, 2014). Additionally, in recent years, logically weaker alternatives to faithfulness have received more attention (Zhang, 2013; Forster et al., 2018). There remain, however, a subset of cancellation cases ("triangle-failures") that are not testable and are just as problematic for these weaker alternatives. Weinberger (2018) and Lin (2019) each propose strategies to justify tentatively ruling out such cancellations without stipulating that they are rare or improbable.

The literature just summarized does not do justice to the most widespread failure of faithfulness: cases involving homeostasis. A simple non-biological example of a homeostasis-producing device is a thermostat, which ensures that the temperature in a room does not vary given would-be causes of random fluctuations. Although such examples predate the formalization of faithfulness (Walters, 1966; Nelson, 2007, pp. 169-171) and are noted throughout the causal literature, discussions of the phenomenon have not causally modeled the dynamical feedback loops by which systems maintain homeostasis. The issue here is not primarily that most graphical causal models do not allow for causal cycles (*pace* Boone (2021)), but rather that they abstract away from core features of the represented systems' temporal dynamics. In contrast, such cases are straightforwardly represented using *dynamic causal models* (Iwasaki and Simon, 1994), which enable one to represent the dynamics of systems away from equilibrium and formally derive their causal behaviors at equilibrium (cf. Blom and Mooij, 2023). Although no one has systematically explored the mappings between such dynamic causal models and the probabilistic relationships among their variables, it is nevertheless possible to show that there are variables that are causally connected in certain dynamic models that will not be probabilistically dependent at the equilibrium time scale, and which therefore violate faithfulness.

Since homeostatic systems are widespread, failures of faithfulness will be widespread. This would appear to be a devastating result for graphical causal inference methods. But this turns out not to be the case. This is because even though a dynamic model would exhibit a failure of faithfulness when extended to a system at equilibrium, it does not follow that the equilibrium model for that system itself exhibits a failure of faithfulness. It follows that homeostatic failures of faithfulness need not pose problems for either equilibrium or dynamic models, provided they are applied to appropriate target systems represented at the relevant time scales.

In fact, the widespread but unproblematic failures of faithfulness in homeostatic systems point to an intriguing novel hypothesis that causal inference is possible only *because* of such failures. Going back to Russell (1912) there is a notion that causal concepts cannot be reconciled with the differential equations of fundamental physics, on the grounds that such equations allegedly reveal that events at any point in spacetime counterfactually depend on events at all other points in spacetime (or at least on all those within the same light cone (Field, 2003)). Frisch (2014) points out that whatever merits this argument has for the fundamental level of description, it does not follow for more coarse-grained descriptions of the variables. Dynamic causal models help explain why finer-grained causal connectedness does not imply systematic coarse-grained interdependence. Failures of faithfulness involving homeostasis are scenarios in which the lower-scale causal connections do not result in higher-scale probabilistic dependence. Rather than leading to a problem for causal inference, such failures help explain why causal attributions are possible in coarse-grained representations of systems, even if the variables in the finer-grained representations exhibit high-degrees of mutual causal interdependence.

2 Background: Markov and Faithfulness

Correlation does not imply causation, since two variables can be correlated without one causing the other. It does not follow that there can be correlations without *any* causal explanation. In fact, standard causal inference methods assume that correlations must be causally accounted for. This idea goes back to at least Reichenbach (1956), whose *principle of the common cause* states that if two variables X and Y or correlated, then either (i) X causes Y, (2) Y causes X, or (3) there exists some common cause, Z, of X and Y such that X and Y are probabilistically independent conditional on Z. In addition to the three explanations Reichenbach proposes for the correlation between X and Y, one can further induce a correlation between X and Y by conditioning on one of their common effects. The assumption that a correlation must be causally explained in one of these ways turns out to be a powerful one for linking causes to probabilities.

In this section, I explain how graphical causal models generalize Reichenbach's principle to an assumption called the "causal Markov condition", and how this condition is supplemented by the causal faithfulness condition. Readers who are very familiar with these principles may skip this section. For readers who are not, who nevertheless are not feeling up to delving into graphical terminology, here are the key points. Like Reichenbach's principle, the causal Markov condition licenses an inference from correlation to causation. Logically, this principle entails that if variables are *not* causally related, then they are not correlated. The upshot is that one can take any arbitrary causal hypothesis represented by a causal graph and derive from it which variables would be uncorrelated if that hypothesis were true. This serves as a means for specifying the set of causal hypotheses that are compatible with a given probability distribution, although additional principles are needed to narrow down this set. The causal faithfulness condition is one such principle. Whereas the causal Markov condition claims that correlation implies causation, faithfulness claims that causation implies of correlation. While faithfulness is not universally true, it is a methodological assumption that simplifies causal inference.

Now for the details. Reichenbach's principle invokes three variables. Using graphical causal models (a.k.a. *causal Bayes nets*) one can provide a representation applying to an indefinite number of variables. The causal relationships among a set of variables are typically represented using a *directed acyclic graph* (DAG) in which the nodes are random variables, the directed edges are direct causal relationships, and there are no cycles (no variable causes itself, either directly or indirectly). A *path* between two variables is a set of connected arrows linking those variables (without specifying those arrows' directions). Y is a *descendant* of X if there is a set of arrows from X to Y all going in the same direction (starting at X).

While Reichenbach's principle states that if there is a correlation, then there is some sort of causal connection, for causal inference the contrapositive of this conditional is more useful. That is, the claim

that if there is no causal connection between two variables, then they will be uncorrelated. A more precise characterization of what is means for two variables not to be causally connected is given by *d-separation*:

Definition 1. ***d-separation*** *(Spirtes et al., 2001): A path is d-separated by variable set* $\mathbf{Z}$ *just in case:*

(a) The path contains a triple $i \rightarrow m \rightarrow j$ *or* $i \leftarrow m \rightarrow j$ *such that* m *is in* $\mathbf{Z}$*, or*

(b) The path contains a collider $i \rightarrow m \leftarrow j$ *such that* m *is not in* Z *and no descendant of* m *is in* $\mathbf{Z}$*.*

d-separation is property of paths. Two variables are d-separated by $\mathbf{Z}$ *iff* $\mathbf{Z}$ *blocks all paths between those variables.*

d-separation formalizes the different ways that one can condition on a variable along a path linking two variables in order to make that path not induce any probabilistic dependence between those variables.

The *causal Markov condition* (CMC) states that if two (sets of) variables X and Y are d-separated by vector of variables $\mathbf{Z}$, then X and Y are probabilistically independent conditional on $\mathbf{Z}$. The causal Markov condition is a generalization of Reichenbach's principle. Although it is not entirely uncontroversial, given suitable qualifications it is plausibly a general truth about the world and is indispensable for causal inference. Many of those qualifications involve restrictions on the variable sets to which CMC applies. CMC will not obtain for variable sets omitting common causes of variables in the set or that involve variables that are logically, mereologically, or metaphysically related. Like the principle of the common cause, CMC does not state that conditioning on *any* cause of two variables will render them probabilistically independent, but only that there exists some such cause. Since common causes that are incomplete or non-proximate will not induce independence, alleged counterexamples to CMC involving insufficiently specified common causes will not be genuine counterexamples. More subtantively, there is an unresolved debate about whether CMC applies in quantum mechanical contexts. Here is not the place to rehash this debate, beyond noting that some of the issues that arise concern whether the variables in quantum contexts meet the qualifications just mentioned. Finally, although I have informally talked about correlations, CMC is a relationship between causation and *probabilistic dependence*. Cases in which causally unrelated time-series are in some informal sense "correlated" but not in fact probabilistically dependent are thus not genuine counterexamples (Hoover, 2003).

Given a causal graph, one can use CMC to determine which variables would be probabilistically independent conditional on which others, were that graph to be true. This allows CMC to function as a compatibility criterion for linking causal hypotheses to joint probability distributions over the random variables in the model. Such a criterion is required in order for causal hypotheses to have empirical consequences. Even with CMC, however, additional principles are required to narrow down the set of candidate hypotheses. To see why, note that for any set of variables, the fully-connected graph in which every variable is directly causally linked to every other does not contain any d-separated variables, and thus trivially satisfies CMC. This is so even if the probability distribution contains many conditional independence relations that the fully connected graph does not entail. The causal faithfulness condition rectifies this limitation. It states that if X and Y are *not* d-separated by $\mathbf{Z}$ (i.e. they are *d-connected*), then they are probabilistically dependent conditional on $\mathbf{Z}$. If one assumes both faithfulness and CMC, this yields a 1-1 mapping between the d-separation facts entailed by the graph and the probabilistic independencies obtaining in its probability distribution, thereby significantly narrowing down the set of candidate causal hypotheses.

3 Debates over faithfulness

Independent of whether one is willing to argue that faithfulness is *usually* true or otherwise justifiable, it is clearly not always true. The most widely discussed type of counterexample to faithfulness involves cancelling causal paths. In the model in figure 1, X influences Y indirectly via M as well as directly. The structural equations for the model are as follows:

$$M = \beta X + U_M \tag{1}$$

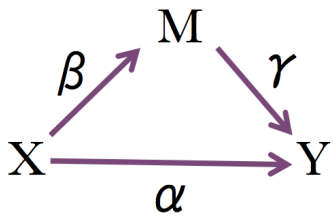

Figure 1: When $\alpha = -\beta\gamma$, there is a triangle-failure of faithfulness

$$Y = \alpha X + \gamma M + U_Y \tag{2}$$

The paths cancel when $\alpha = -\beta\gamma$, resulting in a *triangle-failure of faithfulness*.

The standard defense of faithfulness is that because the set of parameters achieving perfect cancellation are infinitesimally small (Lebesgue measure zero) in relation to all possible combinations of parameter values, it is highly improbable that such failures will occur (Spirtes et al., 2001, p. 41-2). This argument's premise relies on the measure-theoretic fact that imposing the constraint $\alpha = -\beta\gamma$ reduces the parameters' three degrees of freedom to two. This measure-theoretic fact, however, only entails that failures will be improbable given additional assumptions about how to assign probabilities to parameter values (Andersen, 2013). Moreover, there are arguably scenarios in which one should expect there to be cancellation. The two that have received the most attention are what we might call "policymaker counterexamples" and cancellation in homeostatic systems.

Steel (2006) presents the following version of the policymaker counterexample. A city's planners decide to improve the roads, but they worry that this will make people more likely to speed and have fatal accidents. They hire more police officers to counteract this increase, but, for budgetary reasons, want to hire no more officers than those needed to offset the effect of the road improvement. If they succeed, the road improvement will have no net effect on traffic fatalities, even though it is causally relevant to fatalities.

Homeostatic systems are those by which an organism maintains a quantity at a relatively constant value despite external fluctuations. For instance many organisms avoid overheating as the external temperature rises via sweat glands that activate in response to the heat. Intuitively, because the increase in external temperature will have no net effect on internal body temperature, this is a failure of faithfulness. Although some reserve the term homeostatic systems for organisms, it is also common to apply it more generally even to non-biological control systems that are able to self-regulate, and here I adopt this broader terminology. An example of such a control system would be a thermostat that regulates the temperature of a room. Since such systems are designed to ensure that a variable within a system is invariant to external causal influences, they produce failures of faithfulness by design.

One can broadly differentiate policymaker from homeostasis counterexample in that while the former regulate a system via an external mechanism, in the latter the mechanism is internal. But it is not always clear which type of mechanism realistically describes a particular scenario. Although one could imagine a policymaker who builds a model of a system's parameters and gets them to cancel, it is often more plausible that the policymaker would tinker with the system by waiting to see if the quantity of interest deviates from the desired value and then responding when it does. This looks more like the dynamical self-regulation involved in homeostasis. Additionally, consider the opening example in which I exercise with no effect on net calories. One might imagine the scenario as one in which I exercise because I want to eat more without gaining weight, so I carefully coordinate my exercise and eating so that their effects cancel. Alternatively, one could imagine that I just eat when I'm hungry, and that exercise causes me to get hungrier and consume more calories to offset those consumed. While the first description treats me as the policymaker for my own consumption, the second treats my body as a self-regulating device. More generally, policymaker cases can often plausibly be recast as homeostasis examples, so a clearer understanding of the latter will help illuminate both.

To the extent that the literature has explicitly provided a causal model for policymaker and homeostasis counterexamples to faithfulness, it has almost always been with a model like that in figure 1. This is true even for Weinberger (2018), who points out that in order to represent these scenarios, one often needs to include variables that are outside of the triangle (e.g. a common cause for the policymaker's action) and that these additional parts of the causal model are sometimes excluded

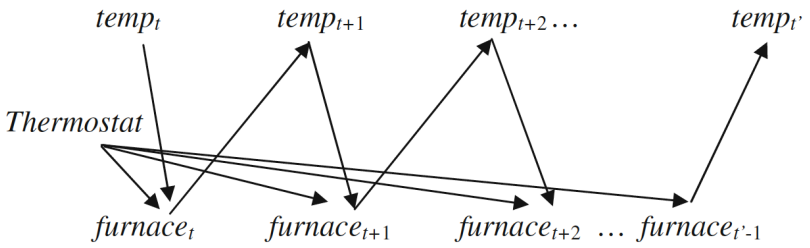

Figure 2: Zhang and Spirtes (2008) thermostat model

by assumptions other than faithfulness. Nevertheless, the cancellation involves a triangle-failure with two cancelling paths. In the following sections, I will argue that homeostasis examples call for an alternative representation.

Not all failures of faithfulness involve cancelling paths. Faithfulness also fails for systems with deterministic relationship as well as examples exhibiting failures of transitivity along a single path (cases in which X causes Y, Y causes Z, $X \rightarrow Y \rightarrow Z$ is the only unblocked path from X to Z, but X is nevertheless not correlated with Z). Nevertheless, the cancellation represented in figure 1 is especially worrisome for causal inference, since it is what Zhang and Spirtes (2008) call an "undetectable" failure of faithfulness. An undetectable failure of faithfulness is a failure of faithfulness in which there exists a false graph that is faithful to the distribution. Given the failure of faithfulness in the figure, the graph $X \rightarrow M \leftarrow Y$ would be faithful to the resulting distribution. In contrast, for the other types of failures mentioned, there will be no graph that is faithful to the distribution. Such failures are "detectable" in the sense that, given weak assumptions, one can determine from the distribution that faithfulness fails. While Zhang and Spirtes (2008) view the fact that faithfulness has testable implications as counting in its favor, Forster et al. (2018) take the fact that there can be no faithful graph to show that faithfulness is too strong. Nevertheless, even weaker principles such as frugality and (P-)minimality will yield the wrong result for the triangle-failure of faithfulness in figure 1, which helps justify the amount of attention such failures have received.

4 Why homeostatic systems call for a dynamical approach

Despite the widespread reference to homeostatic systems in the literature on faithfulness, there are few attempts to causally model them. Notable exceptions are Zhang and Spirtes (2008, p. 251, fig. 4), Weinberger (2018, p. 126, fig. 7) and Boone (2021, fig. 3). Of these, Zhang and Spirtes (2008) is the only one that is explicitly temporal. Their representation of a thermostat is reproduced in figure 2. In it, the thermostat regulates the temperature of a room by adjusting a furnace. Because the temperature after many time steps will be independent of the turning on of the furnace at an earlier time step (due to the proper functioning of the thermostat) there is a failure of faithfulness involving non-transitivity.

Zhang and Spirtes model the thermostat relatively schematically, noting that homeostasis does not involve cancellation (and thus yields a "detectable" failure). The model (figure 2) does not capture how the thermostat functions, beyond its influence on the furnace as a common cause throughout all time-steps. Black-boxing by itself is not a problem, but what's puzzling is how the thermostat can fulfil its role without tracking the temperature, since it must respond to fluctuations. In the model, however, the variable *temperature* is not a cause of *thermostat*, and this cannot be corrected simply by making it a cause, since one would also have to alter whether *thermostat* is a common cause through time. Put differently, although the model involves a cycle between *temperature* and *furnace*, the thermostat is not part of that cycle.

The other models do no better. Weinberger's (2018, p. 124) model of the policymaker objection reflects that a policymaker needs to know the parameter values in the target system to make the paths cancel. But his variable for the policymaker's decision (D in his fig. 4) is just as much a black box as Zhang and Spirtes *thermostat* variable. His dynamical representation of the case (fig. 5) omits any policymaker variable, so does not explicitly represent how the policymaker ensures that the target system reaches its intended state. Finally, his graph of a homeostatic system (Weinberger, 2018, p.126

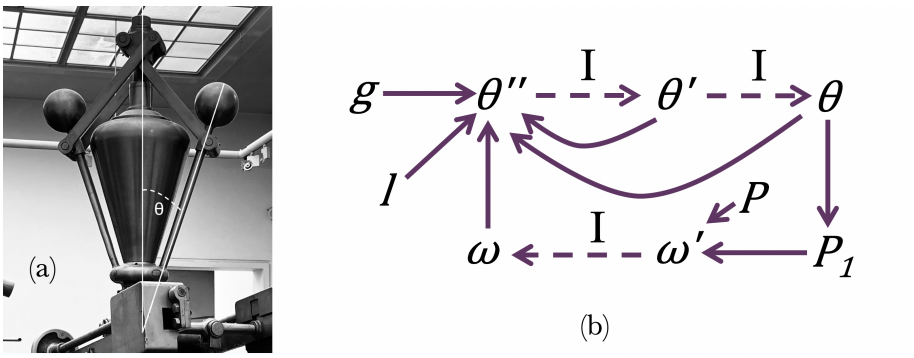

Figure 3: (a) Photo of a Watt Governor indicating angle θ; (taken at the Deutsches Museum) (b) Dynamic Causal Model of the Watt Governor

fig. 7) represents how organisms managing to keep body temperature constant survive (and end up being observed), without describing a mechanism by which they achieve this.

Boone (2021) argues that homeostatic mechanisms are not violations of faithfulness, but rather violations of the assumption that causal graphs are acyclic. This is a plausible proposal, but he does not further specify how to interpret causal cycles in a graph. Additionally, when one moves from acyclic to cyclic graphs, the interpretation of faithfulness changes as well, since d-separation needs to be defined differently (Park and Raskutti, 2016).

Blom and Mooij's (2023) "perfectly adapted dynamical systems" include cases of homeostasis. These authors build on the same dynamic causal modeling tools that I will be relying upon, though their own models differ in a range of subtle ways that I will not catalogue. Mooij's lab is currently the epicenter of work on causally modeling dynamical systems, and merits more philosophical attention than it has thus far received.

5 Dynamically Modeling Homeostasis

The ability of systems to self-regulate using negative feedback loops is well understood within dynamical systems theory. By bridging standard dynamical models and causal models, one can therefore get a much more sophisticated causal understanding of how systems achieve homeostasis. Here I employ Iwasaki and Simon's (1994) modeling framework to develop a dynamic causal model for a paradigm example of a self-regulating device: the Watt governor. Weinberger and Allen (2022) present a preliminary model for the governor. Here I expand their model and provide more rigorous derivation of it. Details regarding the construction of the dynamic causal model from differential equations and the derivation of the equilibrium model through Iwasaki and Simon's equilibration operator can be found in the appendix. Beyond being an example of a homeostasis-producing device, the governor has also been treated as an example of a complex dynamical system whose elements are so inter-related as to not be characterizable as having discrete and manipulable parts (van Gelder, 1995, p. 354). As this might also be taken as a basis for doubting whether it can be causally modelled, the model developed will also illuminate the applicability of causal concepts within complex systems.

The Watt governor (Beltrami, 1987) is a device designed to regulate the speed of steam engines. More specifically, it ensures that the engine rotates at a constant speed despite random changes ("perturbations") to the workload on the system. It does this by creating a negative feedback loop between the speed of the engine (ω) and the amount of steam fed into it. More specifically, the governor is a spindle connected to flyballs via arms such that as the engine spins the flyballs are pushed outwards, changing the arms' vertical angle θ to the spindle (fig. 3(a)). The arms are connected to a sleeve on the central spindle such that as they are pushed outwards, the sleeve is pushed down, and the sleeve is connected to a throttle valve such that as it moves down less steam is fed into the engine, whose speed then decreases. The upshot is that the governor provides an elegant means by which an increase in engine speed above the desired value automatically leads to an adjustment to the throttle valve that counteracts this increase by lowering the speed. It counteracts *decreases* in speed analogously. The engine thereby maintains near constant speed despite external perturbations.

Figure 3(b) presents a dynamic causal model for the Watt governor. This model includes not only θ and ω, but also their higher-order time derivatives: θ' and θ'' for the velocity and acceleration of θ and ω' for the rate-of-change of engine speed (and thus the engine's *acceleration*). The model also includes *gravity* (g) and arm length, (l) as constants, P_1 for the steam fed into the engine, and P for the workload on the system (these labels match Beltrami (1987)). The model contains both ordinary causal arrows as well as dashed arrows with 'I's called integration links. These reflect the mathematical operation of integration (the inverse of differentiation) used to derive the value of a variable at a time step from its next-highest order derivative and its value at the prior time step (e.g. using velocity and position at $t-1$ to derive position at t). We'll see that these links play a subtle role in building time-scale separation into dynamic models.

Let's begin by considering the factors that influence θ''. For purposes of explication, it will be easiest to consider a governor in which the arms open downward, that is, in the opposite orientation as figure 3(a). Gravity g pulls the flyballs downward and the centrifugal force due to the engine rotation ω pushes them horizontally outward. Since the flyballs can only move along a curve determined by arm length l, these forces are constrained to act tangent to that curve (see appendix). The relative contributions of the downward and horizontal forces to the movement of the flyballs along the tangent depends on the angle of the arms, and for a particular engine speed there is a specific angle that the arms will have at equilibrium (when the forces balance). Additionally, θ' influences θ'' via a frictional force that we'll mostly ignore here. In the dynamical model of the governor, the equation for θ'' is identical to that for a rotating pendulum hanging from a hinge.

Whereas the solid causal arrows should be interpreted as linking variables at a time-step, the integration links link variables across time-steps. While g and l influence the current acceleration of θ, it takes some time for θ'' to influence θ's velocity and position. The inclusion of derivative-integration link pairs allows one to causally represent a system at a time scale at which not all variables have had time to reach their steady-state or equilibrium values. Whereas variables linked to their causes via solid causal arrows are assumed to have fully responded to any changes in their causes at the time-scale of interest, the inclusion of a variable's time-derivatives indicates it has not fully responded to changes in its causes at that time-scale, but rather takes multiple time-steps to reach a new steady-state.

To complete our dynamic model, we need to simply note that 1) θ influences ω via the throttle valve, which changes the amount of steam P_1 and 2) that the acceleration of the engine depends on P_1 and the system's workload P. It is significant that ω's influences are represented as causing it via ω'. Because using integration to derive a variable's current value requires its prior value, this value (or at least some initial value ω_0) must be independently specified in the dynamic model. This means that when considering the dependence of θ'' on ω using the pendulum equation, we can treat ω as an exogenously given constant, even though in the broader system it changes as a function of θ. This amounts to treating the rate at which θ responds to changes in ω as very fast compared to the rate at which ω responds to changes in θ.

Iwasaki and Simon (1994) provide an *equilibration-operator*, which, when applied to a variable away from equilibrium, yields a causal model in which that variable has reached equilibrium. θ must be equilibrated prior to ω, since the dynamic feedback loop linking θ to its derivatives must be resolved in order to equilibrate the larger loop going through ω'. Equilibrating θ yields a model in which g, l, and ω directly cause θ, reflecting that θ's equilibrium value depends on these variables (fig. 4(a)). Here equilibration merely collapses the dynamical feedback loop. Equilibrating ω yields more substantial changes (4(b)). The equilbrated model reflects the fact that at equilibrium, the force provided by the steam going into the engine must offset the force due to the workload (otherwise ω' would not equal zero) and this determines the angle of the arms that would allow the required amount of steam, which in turn (along with g and l) fixes the equilibrium engine speed. Readers might be surprised that θ causes ω in the equilibrium model, which is the reverse direction as in the dynamic one. Yet although the value of ω determines θ's value in the short-run, this relationship leaves the long-run value of ω unconstrained. The way the governor system actually governs engine speed is through regulating the steam flow. Arrow reversals such as this have led Dash (2003) to doubt the adequacy of equilibrium models, though Weinberger (2021) defends this consequence of equilibration.

Some may be further surprised that the equilibrium engine speed depends on the workload. Wasn't the whole point of the governor to keep the engine-speed constant? The answer is that although the governor maintains constant engine speed in light of temporary perturbations to the workload, longer-term changes in the workload do lead to a change in engine speed (Beltrami, 1987, p. 167). Even so,

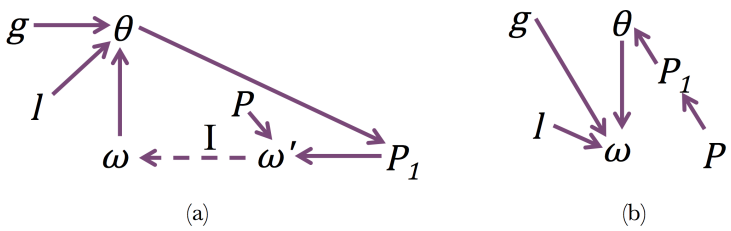

Figure 4: The models resulting from equilibrating (a) θ and then (b) ω in the dynamic causal model. (Derivations in appendix).

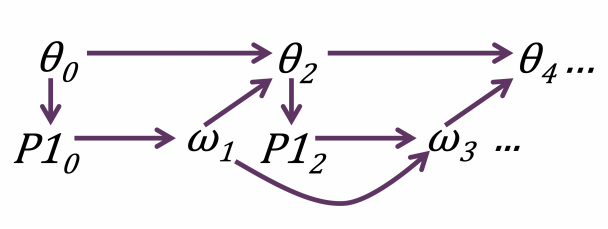

Figure 5: Unfolded Dynamic Model

the governor plays an important role in ensuring the engine has the right amount of steam to do the necessary amount of work.

Although the dynamic causal model contains a cycle via the integration links, the model can be unfolded in time such that there are no cycles. Figure 5 displays how this would be done for a model with just θ, P_1 and ω (P_1 is written as $P1$ so that the subscript not be misinterpreted as a time-step). Variables linked by ordinary causal arrows influence one another within a time step, while variables linked by derivative-integration-link pairs influence one another across time steps. Note that θ also influences itself across time steps, reflecting how unlike equilibrium systems, which are "memoryless", for variables away from equilibrium one needs to know their prior value in order to derive their current one using integration. While the unfolded graph helps illustrate the temporal relationships among the variables, the shorthand dynamic graph in figure 3(b) has the following two virtues. First, it enables one to easily determine whether the system satisfies a necessary condition for stability. The *structural stability principle* (Dash, 2003) states that in order for a variable with a higher-order derivative in the model to have a stable fixed-point, it must be a cause of its highest-order derivative. In our example, this means that both θ and θ' must causally influence θ'', as they do in the graph. Second, the dynamic model allows one to derive the equilibrium model via equilibration.

Iwasaki and Simon's framework is a generalization of Simon's (1953) causal ordering method, which takes a set of equations, and derives a set of structural equations in which each variable is given as a function of its causes. Although this method is not widely known among causation researchers, the structural equations that supplement causal models certainly are, and if one takes a set of equations derived using Simon's method and adds independent error terms to each one, one gets a variable set satisfying the causal Markov condition. Iwasaki and Simon (1994, p. 145) note that Simon's method was originally developed to model static systems in which the variables are measured at equilibrium or steady state. This might seem like a fluke of Simon's framework, but in fact much of contemporary causal modeling that does not explicitly time-index the modeled variables implicitly models the system as static (Weinberger, 2019). The models of homeostatic systems in the faithfulness literature provide a further illustration of how attempts to model dynamical processes using standard causal models do not in fact capture key features of the dynamics. Finally, even if one is not especially interested in the relationship between equilibrium and dynamic causal models, the equilibrium model is what one would get when applying a constraint-based algorithm such as PC to variables sampled at a rate that is longer than it takes for them to reach equilibrium. For all these reasons, dynamic causal models yield insights even for causal modeling applications that do not explicitly discuss a system's dynamics.

6 Failures of faithfulness in homeostatic systems

Now that we have dynamic and equilibrium models for the Watt governor, we can get more precise about the failure of faithfulness involved. Surprisingly, the failure of faithfulness that the governor is designed to achieve – namely, the independence of longer-term engine speed on temporary workload perturbations – does not explicitly appear in either the equilibrium or dynamic causal model. The equilibrium model does not represent any short-term perturbations. The dynamic causal model uses derivative-integration link pairs to represent the feedback loops that bring the perturbed system back to equilibrium, but does not represent short-term changes in exogenous variables such as P. This is true also of the differential equations consulted in building the dynamic model.

Does this mean that homeostatic systems do not produce genuine failures of faithfulness? Not at all. We can still build causal models exhibiting the failure. Consider again the "unfolded" dynamical graph in figure 5. Although the dynamic causal model is intended to represent the system when it is temporarily away from equilibrium, there is nothing to stop one from unfolding that graph for as many time steps as it takes until the system has had time to reach equilibrium. Let's imagine that by $t = 100$ the system will have reached equilibrium. Given some temporary external change to the workload P influencing ω at $t = 0$, one can verify that there will be a causal path from P's value at that time-step to ω_{100}. But, assuming the governor is functioning properly, the variables will be uncorrelated. We thus have a failure of faithfulness.

Homeostatic failures of faithfulness constitute a type of transitivity failure. Yet if one only consults the unfolded version of the dynamic graph, one gains no insight into why the system reaches equilibrium. Only by looking at the dynamic causal model and noting that the system reaches equilibrium if $\theta'' = 0$ does it become clear why a negative feedback loop could achieve this goal. Dynamic causal models therefore illuminate failures of faithfulness in homeostatic systems in a way going beyond standard causal models.

The probabilistic implications of dynamic causal models have yet to be worked out systematically, making it difficult to spell out precise conditions under which there are faithfulness failures. The models developed here are nevertheless illuminating. The fact that the expected failure of faithfulness appears in neither the equilibrium nor dynamic model suggests that it does not pose any problem for those particular models. The unfolded graph does reveal a genuine failure of faithfulness – short-term perturbations will have short-term influences on the system, but the chain of short-term counterfactual dependence relationships will not yield longer-term counterfactual dependence. But the dynamic feedback loops in the dynamic graph clarify how this failure comes about, and the possibility of applying the equilibration operator to get an equilibrium model reveals why such lower-scale failures do not undermine the possibility of causally representing the system at the equilibrium time-scale. In fact, it is the existence of equilibrating feedback loops – along with the failure of faithfulness implied by their actually equilibrating – that gives rise to the substantially different causal relationships in the equilibrium model.

In addition to the failure of faithfulness described, there is an additional failure of faithfulness due to the relationships in the equilibrium model being deterministic. Such failures are rightfully treated as the least worrisome for causal inference, in the sense that modeler will likely notice that there are deterministic relationships in the data. Nevertheless, a non-deterministic example would be required to employ algorithms that infer the causal structure from conditional independencies.

Failures of faithfulness in homeostatic systems do not resemble failures involving cancellation. One consequence of this is that, like other failures of transitivity that Zhang and Spirtes (2008) discuss, the "measure zero" argument that cancelling parameters are unlikely does not apply. In fact, we do not need a system as sophisticated as the governor to show that such failures will be widespread. A simple damped pendulum will tend to return to equilibrium despite perturbations, so its longer-term position will be invariant to causes of perturbations. The governor model reveals how this feature of the pendulum is not limited to small and self-contained feedback loops, and that in fact the feedback loops in smaller subsystems can be built upon to produce new faithfulness failures. The governor exploits the equilibrium-restoring tendency of the pendulum to design a wider feedback-loop for regulating engine speed. While biological homeostatic mechanisms will be even more complex, the governor model has enough complexity to serve as a starting point for thinking about them.

In cases of cancellation, there is no question that the uncorrelated variables are in fact causally related. It is simply that the cancellation makes it harder to discover the causal relationship. In contrast, in homeostatic systems there is a real sense in which the faithfulness violating causal rela-

tionships do not obtain at the relevant time scale. Turning on an oven in a thermostat-regulated room is not a cause of the room's longer-run temperature, and temporary perturbations in workload are not causes of the governor-regulated steam engine's long-run speed. One might insist that causation must be transitive and that therefore such events must be counted as causes. I would argue, however, that such an insistence comes at the expense of appreciating how systems are able to exhibit markedly different causal patterns at different time scales and thus of how lower-scale complexity can give rise to higher-scale simplicity.

For all that homeostatic failures of faithfulness are widespread, they do not pose any evident problem for causal inference. Even when consulting the unfolded version of the dynamic graph to illustrate the failure, there was only a failure of faithfulness when we extended it to the equilibrium time scale. As long as one applies dynamic models to time scales at which a system has not yet returned to equilibrium and equilibrium model to systems at equilibrium, they need not have any false consequences.

I noted earlier that policymaker counterexamples to faithfulness may involve either cancellation or feedback loops. They potentially involve both. The tendency to model them as cancellation alone results from the use of causal models with time-invariant parameters. In Steel's (2006) road improvement example, for instance, the effect of the road improvement on speeding implicitly depends on the number of police hired, though the effect conditional on police is time-invariant. On this picture we imagine the policymaker's hiring of the police as akin to adjusting a dial to ensure cancelling paths, where the policymaker also has knowledge of the time-invariant mechanisms along other paths. If we instead imagine the policymaker as observing traffic fatalities and hiring police in response to any increase, then traffic fatalities influence *police* via its time-derivative *police′* (indicating a lag in hiring). The resulting temporal variation in the effect of road improvements on speeding may help explain why Pearl (2009, p.63) suggests that faithfulness failures are less problematic in longitudinal studies, since at least some of the time there will not be cancellation. We see that even policymaker examples may benefit from more attention to temporal dynamics, although a fuller analysis will have to wait for a separate paper. One intriguing possibility is that if one imagines the feedback loop influencing hiring as happening automatically without a lag at a particular time-scale, this would undermine the possibility of representing the effect of roads on speeding as an invariant parameter that can be exploited for intervention. The concern resembles that raised by the *Lucas critique* in economics.

7 The faithless foundations of causal inference

Causal modeling methods are sometimes viewed as being of limited use for developing metaphysical analyses of causation. Yet, as Weinberger et al. (2023) argue, the suitability of causal methods depends on objective features of the world (the "worldly infrastructure"), and one can thus gain insights into these features by studying causal methods. The features they highlight all concern ways in which a system's variables or relationships can be modeled as being independent of one another, including: the probabilistic independence of the causal Markov condition, more abstract modularity or mechanism-independence principles, near-decomposability, and level-independence. The notion that causal relationships exploit independence relations in the world is held not just by causation enthusiasts, but is just as well reflected in skeptical arguments claiming that the world is interconnected in ways that undermine the objectivity of causal relationships. These arguments arise most commonly within the philosophy of physics (e.g. Field, 2003) and typically trace back to Russell (1912). Explicit anti-causal arguments (e.g. Wagner, 1999) are less common in the life sciences, though some suggest that highly non-modular systems will be inhospitable to causal reasoning (Mitchell, 2009). Finally, discussions of dynamical systems in cognitive science are not primarily concerned with causation, but often claim that dynamical systems are non-modular in ways that might be taken to undermine their being given causal representations. As noted, van Gelder's (1995) influential discussion of the Watt governor asserts that it eludes any non-arbitrary decomposition into modular components.

The dynamic causal models presented here not only illustrate that at least some complex dynamical systems can be modeled causally, but further, through elucidating homeostatic failures of faithfulness, help explain why higher-scale probabilistic independence is compatible with lower-scale causal dependence. Cases in which a system is maintained at equilibrium via a negative feedback loop are ubiquitous and the independence resulting from such processes is not coincidental but by design. This helps explain why higher-scale causation is not merely compatible with lower-scale complexity,

but in fact arises frequently as a result of the lower-scale causal feedback loops. Far from undermining causal inference, homeostatic failures of faithfulness produce the independence relationships that license causal attributions in coarse-grained systems.

The implications for this for causal skepticism remain unresolved. The skeptic could legitimately point out that the differential equations for the governor already contained a fair amount of time-scale separation via the employment of integration links, thus indicating that, at the modeled time-scale, not every part of the system influences every other part at the same rate and to the same degree. The skeptic might therefore argue that in a *truly* complex dynamical system, or a genuinely fundamental physical description of reality, every part does influence every other part to the same degree (at least within some very extensive region) and thus will not allow for causal attributions. This worry is especially salient in the context of physics, where causal skeptics such as Field (2003) point to the hyperbolic differential equations of general relativity, which define a light cone structure, and are of a different type than those for the governor.

Here my goal is not to refute causal skepticism. Should it turn out that the possibility of making causal attributions is sensitive to the type of differential equations used to model a system, this by itself would be a promising and almost entirely unexplored research topic. Previously, the only one to explore the relevance of differential equation type for causal attributions is Smith (2000), whose analysis diverges from that here in presuming that hyperbolic equations will be *most* amenable to causation. The work coming out of Mooij's lab since Mooij et al. (2013) is rapidly shedding light on the range of dynamical systems that can be causally modeled, though this literature points towards unresolved novel questions about which representations are causal. To give one example, are *causal constraints models* (Blom et al., 2020) genuine causal models? Here is one area where philosophers are well situated to contribute. Whatever the results of this research project, we can already draw an important lesson: since the possibility of making causal attributions is potentially sensitive to distinctions among differential equations, any general statements about causation in complex dynamical systems should be viewed with caution.

8 Conclusion

This is the first paper to suggest that failures of faithfulness can be a good thing for causal inference. This is partly due to the fact that prior causal modeling work has mostly been unable capture the dynamical properties of a modeled system, and partly due to a neglect of how causal representations are relative to scale. Since the empirical content of graphical causal models is given by their probabilistic implications, the scale-relativity of causation requires that the probabilistic relationships will also be scale relative. The models discussed here explain why this will in fact be the case. This explanation presupposes widespread failures of faithfulness, and this is exactly as it should be.

A Building and Equilibrating the Dynamic Model for the Governor

Here is the primary equation for the dynamic model of the governor, describing the behavior of the arms connected to the spindle:

$$(1)\quad \frac{d^2\theta}{dt^2} = \underbrace{(n\omega)^2 cos(\theta) sin(\theta)}_{\text{(i)}} - \underbrace{\frac{g}{l} sin(\theta)}_{\text{(ii)}} - \underbrace{r\frac{d\theta}{dt}}_{\text{(iii)}}$$

Since this equation is equivalent to that for a rotating pendulum hanging from a hinge (Beltrami, 1987, pp. 152-5 example 6.5), it is worth starting by modeling the latter system, without yet thinking about the additional cycle from θ to ω via the throttle valve in the governor. When modeling the motion of the bob hanging on the pendulum, all forces acting on the bob are given by vectors running tangent to the curve traveled by the pendulum (figure 6(a)). Term (i) in the equation indicates that the horizontal centrifugal force due to the rotation of the pendulum ω is directed outward along the tangent (and thus partially upward). Term (ii) in the equation indicates that the downward force due to gravity is directed inward along the tangent. Term (iii) indicates the existence of resistance needed

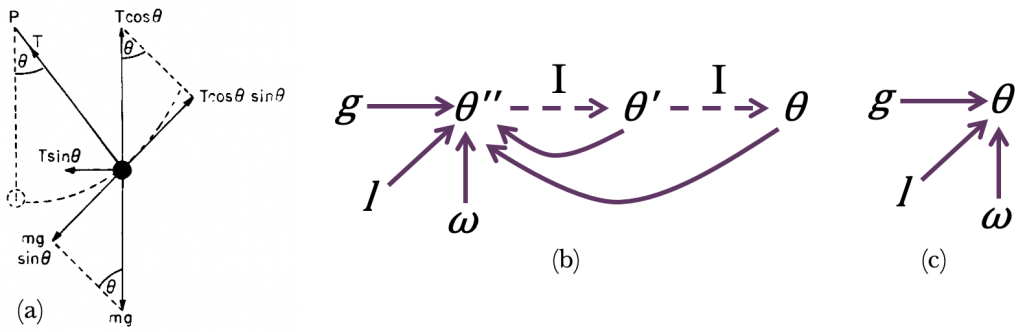

Figure 6: (a) Direction of forces for a rotating pendulum (b) dynamic model (c) equilibrium model from equilibrating θ

to ensure that the pendulum in fact does not move at equilibrium rather than rotating indefinitely, but will not be otherwise important for the analysis that follows.

The dynamic model for this system is given in figure 6(b). In addition to equation (1), one needs to further specify that (A) the variables g, l, and ω are independently given (i.e. are *exogenous*), and (B) that when one derives a variable from it's higher-order derivative via integration, this requires specifying that variable's value at the prior time-step. Concretely, this corresponds to the following equations:

(2) $G = g$
(3) $L = l$
(4) $\Omega = \omega$
(5) $\theta_t = \theta_{t-1} + \theta' \Delta t$
(6) $\theta'_t = \theta'_{t-1} + \theta'' \Delta t$

The integration equations in (5) and (6) are presented as discrete approximations, though could of course be given a continuous version. Because using a derivative to predict a variable's value at a time-step requires that variable's value at the prior time-step, variables with higher-order derivatives in a dynamic model are treated as exogenous within that model. The idea behind Simon's causal ordering method is that facts about what causes what depend on the order in which one can solve for the values of variables in sets of equations such as (1) to (6). Given the value of θ_{t-1} in equation 5 and the values of the variables given by equations (2)-(4), one can use equation (1) to solve for the value of θ''. Accordingly, all four of these variables cause θ'', which is then used to derive θ' and θ at later time-steps via integration.

Dash (2003) provides a schema for deriving the model resulting from equilibrating X:

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 θ requires replacing (1) with:

(1') $0 = (n\omega)^2 cos(\theta) sin(\theta) - \frac{g}{l} sin(\theta)$

Since θ is no longer exogenous and the values of the other variables in (1') still are, θ now depends on those variables, yielding the equilibrium model in figure 6(c).

Now that we understand the part of the governor system corresponding to the flyballs, let's now consider how the angle of the arms regulates the steam flow. The dynamic model is given in figure 3. As described in the main text, θ influences ω' via the flow of steam P_1 and ω is derived from ω' via integration. For the purpose of understanding how the governor regulates the steam flow, we can treat the flyball system as a black box whose input is ω and output is θ. Given a particular value of ω, ω_0, there will be a corresponding value of θ, θ_0.

The description of the system thus far clarifies how the governor is able to return the engine to an equilibrium speed in light of perturbations. This is accomplished via the negative feedback loop involving the throttle valve. Yet this does not explain which factors determine the particular equilibrium value of ω that the governor reaches. This involves some subtle dynamics that I will go into. Before getting there, there is a shortcut that will allow us to complete the causal model. The change in the speed of engine (i.e, its acceleration) is given by the rotational version of Newton's second law:

(7) $I\omega' = P_1 - P$

Here I is the rotational inertia of the flyballs, which we'll ignore for now. Strictly speaking, P_1 and P refer to the *torque* due to the action of the steam and the load on the system, respectively. P is itself an exogenously given constant Since the amount of steam is determined by the angle of the arms, we know that P_1 is a function of θ. Combining these facts, we can add the following equations:

(8) $P = p$
(9) $P_1 = f(\theta)$

Adding equations (7)-(9) allows us to complete the dynamic model in figure 3.

When we equilibrate ω, ω' is set to 0, so (7) and (8) entail that P causes P_1 at equilibrium. From (9) we then get that P_1 causes θ. Note that once we have equilibrated ω, we can no longer treat ω as exogenous, since it no longer depends on its value at the prior time step. We therefore need to return to (1') to once again solve for the ordering. Combined with (2) and (3) plus the derived value of θ, we get the value of ω, yielding the equilibrium model in figure 4.

That concludes the derivation of the causal models in the paper, though once we've come this far it is worth highlighting where the causal discussion in the last two paragraphs diverges from the standard dynamical presentation. Here I've followed Beltrami (1987), which is more or less identical to that in Pontryagin (1962) (the latter attributes the core ideas to Vyshnegradskiy). Confusingly, Beltrami describes P as the "torque due to the variable load on the flywheel" (Beltrami, 1987, p.163). Yet P is itself treated as having a constant longer-term value, and variation in P is only reflected in the influence of this variation on other variables in the system. Specifically, P_1 can itself be decomposed into two components $\bar{P}_1$ indicating what the steam flow would achieve the desired speed and a separate (unlabeled) component indicating the deviation of P_1 from $\bar{P}_1$. This deviation is the result of of variation in the workload. It follows that in equation (7), any temporary changes in the workload are reflected *not* in P but rather in the fact that P_1 is no longer identical to $\bar{P}_1$.

Because a change in the workload will lead to a change in ω and ω influences θ, if the system begins at some value ω_0, with a corresponding θ_0, and then there is a change in workload, θ will have a new, distinct, value. Since both the deviation of the steam flow from its desired value and the deviation of θ from its prior value are proportional to the change in the workload, they are also proportional to one another:

(10) $P_1 - \bar{P}_1 = \alpha(cos\theta - cos\theta_0)$

Here α is a proportionality constant, and is positive. The cosine function derives the height of the flyballs corresponding to a particular angle, where θ_0 indicates the angle before the workload fluctuation and θ the angle after. Combining (7) and (10) yields:

(11) $I\omega' = \alpha cos\theta - F$

Where $F = (P + \alpha cos\theta_0) - \bar{P}_1$. Beltrami then does some additional transformations of the constants to replace I and α, but the upshot is that once one has (11), one can solve for the equilibrium value of θ by setting ω' to 0 and then deriving it from the constants in F.

How should we think of θ_0 and $\bar{P}_1$, which appear equation (11) (as part of F), but not in the equations used to build the dynamic causal model? One might initially be inclined to think of θ_0 as just θ's value when ω is at equilibrium. But ω does not just have a single equilibrium value, but rather different equilibrium values corresponding to different workloads (values of P). Nevertheless, for *any* such equilibrium value, the system will exhibit a resistance to change. Including ω_0 in a model is a way

to model this change-resistance, without privileging any particular value of ω (and certainly without privileging any particular time step).

If there is no longer-term change in the value of P, then the equilibrium value of P_1 will equal $\bar{P}_1$ and the equilibrium value of θ will equal θ_0. If, however, the workload changes, then θ will have some other value at equilibrium. What value this is will depend in part on how far P_1 is from its desired value at the new workload. As with θ_0, the use of $\bar{P}_1$ in calculating ω's value at equilibrium does not reflect that some particular workload is privileged, but rather reflects a more general ability of the system to adjust to different workloads (within a particular range).

How does the causal model manage to represent equilibration without explicitly representing θ_0 or $\bar{P}_1$? When P has a fixed value (i.e. all changes are transitory), then θ_0 just is the equilibrium value of θ for that value of P. When P changes, the change in P corresponds entirely to the difference between $\bar{P}_1$ and P_1. So by considering how the equilibrium values of the variables change given a change in P, one does not need to independently specify that the equilibrium value of P_1 equals $\bar{P}_1$. While the equations in Beltrami's presentation might still be preferable for one just wanting to find the solution to the differential equations, I would argue that the causal model is nevertheless illuminating. The inclusion of θ_0 and $\bar{P}_1$ might misleadingly suggest that the equilibrium state of the system depends on the non-equilibrium values of its variables, or that there's no way to solve the equations without exogenously specifying how much additional steam is needed. In fact, the inclusion of θ_0 reflects that θ is kept stable by the feedback loop in the causal model, and $\bar{P}_1$ is an artifact of the fact that variation in P has not been explicitly represented in the dynamical model.

Acknowledgements

This paper was inspired by Jeff Tweedy's *How to Write One Song*. Thanks to Colin Allen, Mel Andrews, and Nic Fillion for helpful feedback on earlier drafts, John Dougherty for allowing me to repeatedly storm into his office to talk about differential equations, and to several audiences in Munich, including participants at Lauren Ross' Causation and Explanation workshop, for questions and comments.

References

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science 80*(5), 672–683.

Beltrami, E. (1987). *Mathematics for Dynamic Modeling*. Academic Press.

Blom, T., S. Bongers, and J. M. Mooij (2020). Beyond structural causal models: Causal constraints models. In *Uncertainty in Artificial Intelligence*, pp. 585–594. PMLR.

Blom, T. and J. M. Mooij (2023). Causality and independence in perfectly adapted dynamical systems. *Journal of Causal Inference 11*(1), 20210005.

Boone, T. (2021). Robustness and modularity.

Dash, D. (2003). *Caveats for Causal Reasoning with Equilibrium Models*. Ph. D. thesis, University of Pittsburgh.

Field, H. (2003). Causation in a physical world. *Oxford handbook of metaphysics*, 435–60.

Forster, M., G. Raskutti, R. Stern, and N. Weinberger (2018). The frugal inference of causal relations. *The British Journal for the Philosophy of Science*.

Frisch, M. (2014). *Causal reasoning in physics*. Cambridge University Press.

Hoover, K. D. (2003). Nonstationary time series, cointegration, and the principle of the common cause. *British Journal for the Philosophy of Science 54*(4).

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

Lin, H. (2019). The hard problem of theory choice: A case study on causal inference and its faithfulness assumption. *Philosophy of Science 86*(5), 967–980.

Mitchell, S. D. (2009). *Unsimple truths: Science, complexity, and policy*. University of Chicago Press.

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

Nelson, E. (2007). Milton friedman and us monetary history: 1961-2006. *FRB of St. Louis Working Paper No*.

Park, G. and G. Raskutti (2016). Identifiability assumptions and algorithm for directed graphical models with feedback. *arXiv preprint arXiv:1602.04418*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pontryagin, L. S. (1962). *Ordinary Differential Equations: Adiwes International Series in Mathematics*. Addison-Wesley Publishing Company Inc.

Reichenbach, H. (1956). *The direction of time*, Volume 65. Univ of California Press.

Russell, B. (1912). On the notion of cause. In *Proceedings of the Aristotelian society*, Volume 13, pp. 1–26. JSTOR.

Simon, H. (1953). Causal ordering and identifiability. In W. Hood and T. Koopmans (Eds.), *Studies in Econometric Method. New York: John Wiley and Sons*, pp. 49–74. New York: John Wiley and Sons.

Smith, S. R. (2000). Resolving russell's anti-realism about causation: The connection between causation and the functional dependencies of mathematical physics. *The Monist 83*(2), 274–295.

Spirtes, P., C. Glymour, R. Scheines, et al. (2001). Causation, prediction, and search. *MIT Press Books 1*.

Spirtes, P. and J. Zhang (2014). A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption. *Statistical Science*, 662–678.

Steel, D. (2006). Homogeneity, selection, and the faithfulness condition. *Minds and Machines 16*, 303–317.

van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy 92*(7), 345–381.

Wagner, A. (1999). Causality in complex systems. *Biology and Philosophy 14*, 83–101.

Walters, A. A. (1966). Monetary multipliers in the uk 1880-1962. *Oxford Economic Papers 18*(3), 270–283.

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: Cambridge University Press.

Weinberger, N. (2021). Intervening and letting go: On the adequacy of equilibrium causal models. *Erkenntnis*, 1–25.

Weinberger, N. and C. Allen (2022). Static-dynamic hybridity in dynamical models of cognition. *Philosophy of Science 89*(2), 283–301.

Weinberger, N., P. Williams, and J. Woodward (2023). The worldly infrastructure of causation.

Zhang, J. (2013). A comparison of three occam's razors for markovian causal models. *The British journal for the philosophy of science*.

Zhang, J. and P. Spirtes (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines 18*, 239–271.