The Human Face of Causation

1. Introduction

Betrand Russell argued for a complete elimination of the concept of cause from our conception of the world in "On the Notion of Cause". Yet while Russell's wholesale rejection of causation may have fit well with the logical empiricism prominent in the early twentieth century, it has many fewer adherents a century later. Russell's skepticism concerning causation in physics continues to have many followers, but twenty-first century 'neo-Russellians' do not follow Russell in concluding from this that the word 'cause' ought to be eliminated entirely from the philosophical vocabulary. Instead it is widely believed that causal concepts play an important or even ineliminable role in common-sense reasoning and the so-called 'special' or higher sciences, such as biology, even if such notions have no legitimate place in the way physics represents the world.

Often the view that causal notions cannot be eliminated from common sense and the special sciences is supported by appealing to an argument made by Nancy Cartwright in an influential paper (Cartwright 1979). Central to the concept of cause is the contrast between causal relations and mere correlations; and it is this contrast, according to Cartwright, that is needed to account for the distinction between ineffective and ineffective strategies. In order to know whether a certain course of action might be effective in bringing about a desired outcome, it is not enough to know various *correlations* between outcomes of the desired kind and other kinds of events whose occurrence we might take to be under our control. We also need to have *causal* knowledge that, as Cartwright argues, goes beyond what can be captured in terms of mere correlations.

Cartwright's own example is that people who carry a life insurance from a certain company whose customers are primarily educators tend to live longer. Merely being told that such a correlation exists does not yet allow us to determine whether purchasing the life insurance is an effective strategy for increasing one's life expectancy. Rather, one needs to know what causal structure underlies the correlation: we need to know whether purchasing the insurance has an effect on longevity or if, more plausibly, the two factors have common causes, such as the high level of education of the insurance members or their access to good health care. In the latter case purchasing the life insurance would not be an effective strategy for increasing longevity. Knowledge of mere correlations, thus, is not enough to allow us to distinguish effective from ineffective strategies.

Cartwright herself ultimately wishes to conclude that causal structures are grounded in causal capacities, which are part of the inventory of the world. For neo-Russellians, however, acknowledging Cartwright's point about the importance of causal concepts in common-sense reasoning and the special sciences, presents a puzzle: how is it that causal reasoning can play an important role in how we conceive of the world even though the world is fundamentally a-causal? Indeed, according to one prominent neo-Russellian, Hartry Field, "the problem of reconciling Cartwright's points about the need of causation in a theory of effective strategy with Russell's points about the limited role of causation in physics [...] is probably the central problem in the metaphysics of causation." **(reference)** One popular strategy for solving this puzzle is to emphasize putatively pragmatic or perspectival aspects of causal notions. This strategy is pursued in various forms by Jim Woodward, Chris Hitchcock, Huw Price, Richard Healey, John Norton, and Field himself, among others, who argue that one can do justice to Cartwright's insight by emphasizing aspects of causal reasoning that are essentially tied to our perspective as human agents interested in making our way about in the world. Once we recognize the "human face" of causation, as Woodward puts it, we can do justice to the importance of causal representations without having to postulate that causal relations are woven into the fundamental fabric of the world. Indeed, the pragmatic or user-dependent aspects of the notion of cause seem to offer further support for skepticism about causation in physics: it is precisely the fact, that causal notions are essentially tied to our human perspective that make their application in fundamental physics strained.

Since they accept Russell's arguments concerning fundamental physics, any positive account of causation offered by neo-Russellians is an account of causal representations or of causal models and their role in our conception of the world, rather than a contribution to the metaphysics of causation. This is quite explicit in Woodward's interventionist account, in which the relata of the causal relation are variables and their values representing properties of the system modeled. Thus, the question concerning causation in physics asked by Woodward and others is whether causal notions can play a legitimate role in how mature physical theories represent the world. This is a perspective I share with the neo-Russellians. Our disagreement is, at least in the first instance, not one over the metaphysics of causation, but over the extent of the domain of applicability of causal models.

Woodward argues that causal notions may cesase to apply in fundamental physics, or that at least their use becomes increasingly strained as we move to more fundamental theories of physics:

> The causal notions and assumptions that figure in common sense do not always transfer smoothly or in an unproblematic way to all of the contexts in which fundamental physical theories are applied. [...] There is thus a (partial) mismatch or failure of fit between, on the one hand, the way we think about and apply causal notion in the upper-level sciences and common sense and, on the other, the content of fundamental physical theories. [...]
>
> My general stance is pragmatic: the legitimacy of causal notions in the upper level sciences is not undermined by the disappearance (or non-applicability) of some aspects of these notions in fundamental physics. (67)

There is a sense in which I agree with Woodward: common sense causal notions do not always transfer smoothly to fundamental physics. Thus, my aim here is not to defend the *full* applicability of common sense causal reasoning to physics or that causal notions in physics must "possess all the features of the notion that figures in common sense and the special sciences." (Woodward 2007, 70) Nevertheless, some of the features that Woodward takes to be paradigmatic of causal representations also characterize the kind of representations constructed with the help of fundamental physical laws. While Woodward has indeed identified some dimensions along which the treatment of at least some systems of physics differ to some extent from that of systems in the special sciences, the differences are less stark than Woodward's discussion suggest and do not warrant the conclusion that causal notions disappear from, or are inapplicable in fundamental physics, or so I shall argue.

There is another point on which I agree with the neo-Russellians. There is a tradition that argues that the truth of all causal claims is ultimately 'grounded' in fundamental physical laws that are causal in character. Fundamental physical laws, according to this tradition, supply a causal foundation for all causal claims, including those in common sense and the special sciences. Woodward finds this sort of "causal foundationalism or fundamentalism" "deeply problematic" (70) and I also agree with Woodward in this respect. In fact, I agree with Woodward's criticism *both* of the view that the legitimacy of causal notions in the special sciences "must somehow reflect or derive from features of causal thinking that can be found in fundamental physics" (66) *and* of the view that causal claims have to be grounded in fundamental causal laws. Yet, as I argued in chapter 1, it is important to distinguish between the content of our physical theories' basic equations and the content of physics: it does *not* follow from the view that causal notions cannot be grounded in the dynamical laws of our physical theories that such notions cannot play a legitimate role in reasoning in physics and that causal claims cannot be part of the content of physics. The content of

physics is not exhausted by the content of our theories' basic equations. That the latter do not provide a causal foundation for the truth of all causal claims does on its own not imply a "mismatch," a "failure of fit," an "unimportance or disutility" of causal notions in physics (see Woodward 2007, 67). Thus, in what follows I am interested in asking what arguments, if any, there are for the stronger conclusion—the conclusion that causal notions can play no legitimate role in physics, and not merely whether causal notions cannot be grounded in ourtheories' fundamental equations.[1]

In the following sections I want to examine arguments by Field, Woodward, and Hitchcock for this stronger conclusion, focusing in particular on Woodward's discussion in his "Causation with a Human Face" (Woodward 2007). I will argue that Woodward's and the neo-Russellians' portrait of causation resembles the face of scientific representation much more generally and, thus, that their arguments fail to establish a genuine contrast between reasoning in physics and the special sciences. Causal representations have a human face, but so do scientific representations in general. The solution to Field's problem concerning the place of causation that I will offer instead is to reject the neo-Russellians' claim that causal reasoning plays no role in our mature theories of physics. Woodward insists that "causal notions are legitimate in any context in which we can explain why they are useful, what work they are doing, and how their application is controlled by evidence" (67). My argument in this book is that causal notions appear in physics in just such contexts.

Most of the arguments I will examine in this chapter have the following form: There is a certain feature *x* that is characteristic of causal models. Representations in physics do not (or even could not) have feature *x*. Therefore, causal notions can play no role in physics. Some of the arguments I will consider appeal directly to an interventionist account of causation, arguing that the notion of intervention that is at the heart of his proposal for elucidating causal notions cannot be applied in fundamental physics. Other arguments appeal to somewhat broader putative characteristics of causal relations and do not presuppose an interventionist account. The features I will discuss in section two are that causal relations are coarse-grained and that all causal claims are claims relative to certain background conditions. My reply here will partly be to appeal to my discussion in the previous chapter and argue that the features of causal representations to which

---

[1] While Woodward is unequivocal in his rejection of causal foundationalism, he is rather more tentative in extending the reach of his arguments to the place of causation in physics more generally. Since my ultimate aim is to examine the strength of various anti-causal arguments, I shall, for ease of exposition, gloss over some of the hesitation Woodward expresses. My argument, then, is with the neo-Russellian inclinations Woodward seems to be drawn toward, even though in the end he might embrace them only tentatively.

Woodward and others draw our attention are features of scientific representations much more generally. In sections three through five I will examine arguments appealing to connections between the notion of causation and of intervention.

One more preliminary remark. Woodward asks if causal notions are applicable to fundamental physics and, as an example of a fundamental theory, cites classical electrodynamics. Thus, I take him to mean by 'fundamental theory' what I refer to as a 'mature theory' of physics. Thus, in what follows I will be using the two terms interchangeably. "Fundamental physics" as I will be using the term, following Woodward, does not merely refer to general relativity, quantum field theory, or perhaps a 'final theory' of quantum gravity, but includes theories like classical mechanics or classical electrodynamics as well.

2. Coarse-Graining and Background Conditions

Paradigmatic common-sense causal generalizations are claims like 'striking a match causes it to light'. That is, they are claims relating relatively small numbers of events or event-types; and the events at issue are relatively coarse-grained macro-events. Implicit in common-sense causal claims is further a distinction between causes and background conditions. Intuitively, the striking of the match appears to be a cause of the flame, while the presence of oxygen is not, even though both conditions are necessary for the match to light. Coarse-graining and a distinction between causes and background conditions also seems to be characteristic of causal claims in the special sciences. This, according to Woodward and Field, contrasts sharply with the case of physics, where the dynamical laws require precisely specified 'initial' states as input, defined over a global time-slice of the world or at least over an entire cross-section of the past lightcone of an event to be determined, allowing no distinction between background conditions and dynamically relevant factors: "In contrast to the incomplete relationships of limited invariance between coarse-grained factors that are characteristic of the upper level sciences, fundamental laws typically take the form of differential equations, deterministically relating quantities and their space and time derivatives at single spatiotemporal locations" (Woodward 2007, 83). What is more, the relations between different spatiotemporal locations are such that "*information about what happens at an earlier time can't suffice to determine the event unless it includes information about each point at that time that is within the past light cone,*" as Field emphasizes (italics in original). Thus, it may seem that there are putatively essential characteristics of causal relations—that they relate a small number of coarse-grained and imprecisely defined localized events, relative to certain background assumptions, to one another—and that these characteristics are absent in representations in physics. There are

several distinct putative contrasts to which Woodward and Field appeal and which I want to examine separately, beginning with the issue of coarse-graining.

Woodward's claim that the "variables of upper level causal theories are extremely coarse-grained from the point of view of fundamental physics" (80) suggests the following argument:

2.1. It is an essential component of causal notions that they relate a small number of coarse-grained and imprecisely defined events to one another.

2.2. The dynamical models of our well-established theories relate to one another states that are precisely defined over entire time slices.

2.3. If causal relations play a legitimate role in well-established theories of physics, then their relata are input states into a theory's dynamical laws.

2.4. Causal notions do not play a legitimate role in any well-established theories of physics.

How good is this argument? Premise 2.1 and 2.2 state that there is a contrast between imprecisely defined and coarse-grained causal relata, on the one hand, and the input state of dynamical laws characterized in terms of the exact values of certain variables, on the other. Yet this contrast is in danger of confusing relations between the variables within a theory's models with the relation between a model and the physical systems it is meant to model. Both in the case of putatively causal claims made by a scientific theory and in the case of representing a phenomenon with the help of a theory's dynamical laws, we need to distinguish carefully the *models* constructed with the help of the theory from *the real world systems* the models are intended to represent. *Within* a dynamical model, a theory's basic equations relate states precisely defined in terms of the values of certain state variables to one another. But within a causal model, the causal relations between variables can be similarly precisely defined, as the formal approaches in (Pearl 2000) or (Spirtes et al) can show. Thus, both causal and purely dynamical models can contain formally precise relations between precisely defined states of the model.

By contrast, in the case of both causal models and purely dynamical state-space models *the relation between a model and the real world system* it is meant to represent is to some degree vague, imprecise, and approximate. All physical theories, as I have argued in chapter 2, represent the phenomena in their domain only within certain limits of accuracy and with the help of abstractions and idealizations: neither causal nor purely dynamical modeling provides us with "perfect models" of the phenomena.

Finally, we can ask about the relations between models constructed with the help of different theories at different levels of grain—and this is the relation on which Field and Woodward mainly want to focus. Woodward maintains, crediting Chris Hitchcock for this point, that in the

case of causal models the "coarse grained variables may fail to completely partition the full possibility space from the point of view of an underlying fine-grained micro theory." (Woodward 2007, 81) What Woodward and Hitchcock have in mind here is the fact that from the perspective of an underlying micro-theory there may be states that are intermediate between two distinct macro states such that it may be unclear or vague which macro-state these intermediate states realize or instantiate. Field argues that causal variables are coarse-grained in a similar sense. He takes it to be "fairly significant" that causal variables are inexact and defines that "a variable is inexact if the claim that it assumes a given value on an occasion can be realized in many different ways that on a deeper level of analysis are importantly different." **(Field)**

But again this feature does not constitute a difference between causal models and the dynamical models of a physical theory since the latter also fail to completely partition the possibility space of any underlying even more fine-grained theory, at least in the case of all but the most fundamental theory of physics—such as perhaps a final theory of quantum gravity. Consider one of Woodward's own examples of a putatively fundamental theory—classical electromagnetism. Like models in the special sciences, this theory fails to partition the full possibility space from the point of view of an underlying fine-grained theory. Even though specifying the state of an electromagnetic system may require specifying the charge densities and field strengths at each spacetime point within some region, possible classical states do not completely partition the possibility space of an underlying quantum field theory, which, for example, permits of superposition states that have no classical analogue.

Thus, as far as the issue of *coarse-graining* is concerned, there is no difference in kind between causal models and purely dynamical physical models—not for relations within a model, nor for model-world relations, nor for inter-theoretic relations is. Moreover, it is unclear why variables would have to be inexact or coarse-grained in any of these senses in order to be causally related. That is, it is unclear why a causal representation would have to be coarse-grained *qua* being causal.

Both Field and Woodward sketch an argument for this claim, suggesting that certain asymmetries among correlations that can exist among the variables in causal structures crucially depend on the fact that these variables are coarse-grained from the perspective of a second more fine-grained. Consider a so-called 'screening-off condition', according to which joint effects E_1 and E_2 of a common cause C, for which $Pr(E_1/E_2) > Pr(E_1)$, are rendered probabilistically independent by conditionalizing the effects on the occurrence of C: $Pr(E_1/E_2\&C) = Pr(E_1/C)$. One can show that, assuming determinism, there will also be some event C^* in the future of E_1 and E_2 that will screen

off E_1 from E_2, threatening the idea that common cause structures exhibit a characteristic asymmetry. As Field and Woodward point out, the asymmetry can only be preserved if we add certain restrictions on the choice of variables, such as that the variables in question represent spatio-temporally localized events. The future screening-off condition C^* will in general not represent a localized event, and hence can be excluded from the model. But contrary to Field's suggestion that what is at issue here is whether or not the variables in question are exact, and that "with 'exact' variables [...] the asymmetry completely disappears in classical physics" (Field,) it is not true that the asymmetry depends on the assumption that the variables in question are coarse-grained in the sense of being multiply realizable by variables on a more fine-grained level. If anything, the opposite is true. *At a given level of grain* at which there are deterministic laws of evolution, there will be a set of natural or salient localized variables that is 'causally privileged' over a set of highly non-natural, non-localized variables, even though the two sets of variables may be interdefinable. But this has nothing to do with the question whether the variables at issue are exact in Field's sense: both 'natural' and 'unnatural' variables are introduced at the same level of grain and are coarse-grained from the perspective of a more fine-grained micro-theory. Moreover, moving to a more-fine-grained level does not make the asymmetry disappear, as long as we make sure that the more fine-grained variables are similarly 'natural'. Yet if we assume that in a certain model correlations between pars of variables are screened off by the value of some third value, it does not follow that there will also be a screener-off for all correlations, if we move to a *more* coarse-grained state space, as Frank Arntzenius has argued (see Arntzenius 1993): if we allow that macroscopic events can have microscopic causes, then a screening-off condition that holds at a certain level of grain can fail to hold for more coarse-grained models.[2] (I will return to this issue in in chapter four.)

A second putative contrast between causal models and models in fundamental physics concerns the role of *background conditions*. Common sense causal regularities, such as 'striking a match causes the match to light' relating relatively small numbers of events or variables to one another, hold only if certain additional conditions apply, but these conditions have the status of background conditions and are not part of the causal regularity itself. By contrast, the models of fundamental physics, according to Field and Woodward, provide us with complete representations of a phenomenon that include all possible influences on the phenomenon in question. Thus, within

[2] The reason for this is, as Arntzenius points out, is that the common cause principle cannot hold for a class of events that has causes outside that class. Arntzenius concludes from this, that the only class of events for which the common cause principle could possibly hold is the entire class of microscopic events. (Arntzenius 1993, 230).

the context of a complete micro-physical theory of the match and its surroundings, it would seem, the complete state of a region of the world prior to the lighting would have to be specified in order to be able to derive that the match will light. Hence we would no longer be able to draw a distinction between causes and background conditions and not only the presence of oxygen but, more worrisome perhaps, also the micro-state of the room next door would come out as a cause of the lighting of the match just as the state of the match during its striking will. In fact, assuming a relativistic theory the entire backward light cone of an event *E* will come out as a cause of *E*. But this, it may seem, is an absurd conclusion—clearly (one might say) the striking of the match is a cause of the lighting in a way in which the state of the room next door is not—and in order to resist it we may want to conclude that the concept of cause is essentially a concept meant to express the kind of dependencies between small numbers of variables relative to background conditions, with which the special sciences or common sense are concerned.

Perhaps, then, the crucial contrast between causal models and dynamical models is not related to the question of imprecision or *coarse-graining* (where there is no contrast) but rather involves the issue of *completeness*. If models require complete cross sections of past lightcones as input, then we cannot draw a distinction between causes and background conditions (such as the distinction between the roles played by the presence of oxygen and the lighting of a match) and we seem to be required to include factors in the model that we would wish to leave out of consideration completely, such as the state of the room next door as we light the match. This suggests the following argument:

2.5. Physical theories provide us with complete models of the phenomena in their domain, constructed from the data on complete lightcone cross sections as input, which do not permit a distinction between causes and background conditions.

2.6. If a model does not permit a distinction between causes and background conditions, then the model cannot be interpreted causally.

2.7 Therefore, there is no place for causal notions in physical theories.

One response to this argument is to challenge 2.6 and argue that what we learn from models of relativistic theories which take into account entire cross sections of backward light cones is that an event has many more causes than we might have naïvely assumed. "What's the big deal in that?" Field himself asks rhetorically, after presenting a version of this argument. Yet he immediately replies that this response is unsatisfactory,

> since there would be a big deal if we had to conclude that if *c*1 and *c*2 are both in the past light cone of *e* then there is no way of regarding one of them as any more a cause of *e* than the other: then Sam's praying that the fire would go out would be no less a cause than Sara's aiming the water-hose at it, and the notion of causation would lose its whole point. (Field 2003, 439)

Accepting all events in the backward lightcone of an event as cause, Field maintains, would eviscerate our notion of cause—a conclusion that could only be avoided, according to Field, if we give up the idea that only concrete events can be causes.

I will examine Field's claim below, but first I want to focus on premise 2.5 in the argument. As I stated it, the argument is not valid, at least on one reading of 2.5. In order to render the argument valid 2.5 has to be replaced or disambiguated by the following condition:

> 2.5* *All* representations with which a physical theory provides us are complete models of the phenomena in their domain.

If not all models in physics were complete lightcone models, then we could at most establish that such complete models could not be interpreted causally, but not that there is no room for causal notions in physics in general. But 2.5* implies (at best) an overly narrow conception of how physical theories represent the world. Field suggests that it follows from the fact that "physics has progressed by replacing [causal claims involving a small number of variables] by differential equation" that modeling in physics in general needs to take into complete initial states, involving a complete specification of an entire cross section of the backward light cone of whatever event we are interested in, yet this is wrong for at least the following two reasons.

First, characterized at the most general level, the procedure for solving the dynamical equations for a system involves specifying *both* initial conditions *and* boundary conditions. Initial (or final) conditions specify the state of the system at one time—that is on one time-slice or on a spacelike hypersurface. Initial conditions are *temporal* 'boundary conditions'. Proper boundary conditions, by contrast, specify the *spatial* boundaries of the system under investigation across time. Such boundaries might be the state of the walls of a container, for example, or, in the case of a mechanical system, the satisfaction of certain constraints, but they need not correspond to the state of some material object through time. A solution of the dynamical equations represents how a system, with specific initial and boundary conditions evolves in time. **[see figure 1a]**

In the case imagined by Field, that of a relativistic system with initial conditions specified on the entire cross section of the backward lightcone of the spatial region occupied by the system at the time *t* of interest, there is no need to specify spatial boundary conditions. **[see figure 1b]** If no

influence can travel faster than the speed of light, then nothing occurring outside of the past lightcone can affect the state of the system at *t*. But the vast majority of models of actual physical systems are not constructed with the help of a pure relativistic initial value problem and need to include a specification of boundary conditions as well. After all, if we wanted to model the evolution of a system during a time period even as short as one second, the cross section of the backward lightcone on which we would have to specify the relevant initial data would have a diameter of 300,000km. Thus, in most actual applications of physical theories systems are treated by specifying the values of variables only within a spatial region much smaller than an entire cross section of the backward light cone and the evolution of the system is determined under the assumption of both initial and boundary conditions.

Caption: While Field's argument presupposes that all models of physical systems, at least those constructed in the context of our mature theories are of the pure-relativistic-initial-conditions-form, depicted in **[figure 1b]**, the vast majority of models used to represent actual systems are models involving assumptions of both initial and boundary conditions, as depicted in **[figure 1a]**.

How does the existence of models with boundary conditions affect the argument above? Arguably boundary condition can play a role analogous to that of background conditions in causal models in the higher sciences: an arrangement of gears in a mechanical model provides the background conditions under which certain dynamical relations hold, just as the presence of oxygen is a background condition for the lighting of the flame. And that certain quantities are zero on a spatial boundary is a shielding condition similar to one that allows us to exclude the state of the room next door in evaluating the causal claim that striking the match causes a flame.

2.5* is false for a second reason. Very often, even within the context of arguably 'fundamental' theories, we represent phenomena not in terms of a complete initial- and boundary value problem, which would provide us with information about the precise physical state of each spacetime point in the spacetime region of interest (as Field demands), but by specifying only a finite number of relevant components of a system at diferent levels of grain and then showing how these components interact. While the use of differential equations might constitute progress, this this has not led to physicists to replace incomplete and partial models with the complete light-cone models, as Field suggests. I argued in chapter 2 that *all* scientific representations are partial and

distort, but for present purposes it is enough to show that there are *some* models constructed within the context of mature or fundamental physical theories that are incomplete and partial.

As a case study, to which I will return repeatedly in this chapter, I want to discuss how the proton beam in the Large Hadron Collider (LHC) at CERN is modeled **(see, e.g., Steinhagen; LHC study group)**. I choose this particular example, since it is as good an example as any of modeling in physics involving fundamental theories, in this case Classical Electrodynamics, Quantum Mechanics, and, of course, the Standard Model of particle physics, which unifies the electromagnetic, weak, and strong nuclear interactions. The LHC is a circular accelerator in which two proton beams are accelerated and then brought to collide with each other. Some components essential to its operation are bending magnets which are responsible for keeping the beams on their circular orbit around the LHC ring, focusing magnets that prevent the beams from spreading and focus the particle orbits at the center of the LHC vacuum chamber, and accelerating systems accelerating the protons. Measurements of the collision products are intended to provide useful tests of the Standard Model.

What is important for our purposes here is that the proton beams in the accelerator and the decay products in the particle detectors are not represented in terms of a complete quantum field theoretic model—that is a model including the quantum field theoretic micro-state of the world in an entire cross section of the backward light cone of the protons' orbits. Rather, the model is partial, ignoring many influences on the accelerator and detector that should in principle be present and treats different components of the experiment in terms of different theories at different levels of grain. Interactions between the proton beam and the accelerator are treated largely using the resources of classical electrodynamics—that is, a theory of fundamental physics, but one that is considerably more 'coarse-grained' than the Standard Model itself. The proton beam is taken to interact directly only with the highly localized electromagnetic fields produced in the various components of the accelerator, such as bending and focusing magnets. The different pieces of hardware that comprise the LHC are modeled macroscopically—the precise states of all the subatomic particles that make up, for example, the bending magnets is not part of the model and there is no full quantum-field theoretic model of the detection chambers. Similarly, the state of the collision products in the particle detectors is not modeled by specifying the full quantum-field theoretic state on a cross section of the backward light cone of the decay products and their environment. External environmental influences on the accelerator are also modeled, both qualitatively and quantitatively, but physicists are selective in deciding which external perturbations need to be included in a model of the proton beam and all such perturbations are

modeled at an extremely coarse-grained level from the perspective of the more fundamental theories at issue. For example, motions of the ground surrounding the accelerator are included in the model by representing the accelerator as being embedded in a homogeneous viscous medium[3] that exhibits both random and coherent motion, which is modeled as plane waves propagating in the medium. These models are compared with seismic measurements. The effects of tidal forces on the accelerator geometry are modeled as well, yet the micro-state of a pot of cheese fondue in a nearby mountain restaurant is not included in the model.

Thus, even the models that physicists construct to search for the remaining missing piece in the confirmation of the Standard Model—the Higgs boson—are partial, involve different levels of coarse-graining and a distinction between relevant factors and background conditions. Of course, there are scientifically legitimate reasons for this treatment: The proton beam is sufficiently well-shielded from may external influences that these can be ignored in the model; and the precise state of the atoms in the bending magnets or of the surrounding mountains does not (within limits) appreciably affect the total magnetic field and hence can be treated in a coarse-grained manner. There are, then, good reasons for ignoring many perhaps in principle existing influences on the protons' trajectories and on the collision products; and there are good scientific reasons for treating some of the influences that do exist in what from the perspective of the standard model is an extremely coarse-grained way. But of course the very same kind of reasons allow us to use partial and coarse-graining representations in the construction of paradigmatically causal models in the special sciences or in common-sense. The state of the neighboring room can be ignored in our causal model of the lighting of the match, since the wall between the rooms sufficiently shields the match from any potential influences on the lighting resulting from changes to the state of the neighboring room. And the precise microscopic state of the match (within limits) will not affect whether or not its lights.

Thus, again, we have not been able to identify a genuine difference between modeling in physics and causal modeling elsewhere. Modeling in fundamental physics, too, proceeds by representing explicitly only a limited range of factors in a model, some of the included factors are treated at different levels of grain, while other factors that possibly might have an influence but are in the circumstances at issue well shielded are relegated to the status of only vaguely specified

---

[3] Steinhagen explicitly discusses how this assumption distorts: "For the application in the SPS and LEP/LHC model, one can assume that the accelerator tunnels are entirely embedded in a homogeneous ground. While this is certainly true for the SPS, the assumption holds for the LEP/LHC tunnel as well, since only about 3km of the 27km circumference is located in hard rock regions of the Jura mountains." (68)

background conditions. In both dynamical models and causal models, background factors do not show up within the model but are appealed to as part of the justification why a given model is appropriate in the circumstances at issue.

As somewhat of an aside as far as the present argument is concerned, it is worth pointing out that the ways in which the functioning of the LHC is described even in the scientific literature, supports Cartwright's thesis that even at the level of fundamental research in physics, our conception of the world is ineliminably causal. For example, in a report from the LHC study group of CERN on the LHC design (LHC Study Group 1995), we learn that there are various places in the machine where beams can be "injected", other components allow to "suppress" dispersion, while others allow for the "clean up" of the beam. Finally, there is the "beam dump" where the beam can be deposited with the help of "kickers". (Thus, the report informs us that "the principle of the beam dumping system is to horizontally kick the circulating beam into an iron septum magnet.") All these words describe what Cartwright would take to be 'concretely fitted out' instances of causings and, hence, one gets the impression that the operation of the LHC, from the injection of proton beams to their collision in the detectors or their being dumped in the beam dump, is an inherently causal process.

Field maintains that the problem with putatively causal models of a physical phenomenon would be that "facts about each part of the past light cone of an event are among the causes of the event." But the models physicists in fact *use* to represent actual phenomena very rarely if ever are the complete structures which philosophers postulate, since they are in general much too complex for us ever actually to possess.

How can it be that physicists manage to construct such partial and idealized models given the purported feature of differential equations to which Field draws our attention? How can it be that models in physics are partial, if a differential equation requires specification of the precise data for each spacetime point on an entire initial value surface as input? A discussion of the various strategies by which actual models avoid the unwieldy complexities of complete initial value problems over cross sections of entire lightcones would take us too far afield; but among the strategies are the following. First, as we have seen, models may include boundary conditions that represent facts about shielding conditions and which allow us to severely restrict the spatiotemporal region which has to be represented in a model. Without such a restriction, representing actual phenomena and confirming the adequacy of our representations would be simply impossible.

Second, models may set the values of variables identically equal to zero in large subregions of the initial value surface or may not even contain variables representing possible influences arising from large regions of the initial value surface, if whatever is physically going in these regions can be taken to be irrelevant to the phenomenon at issue. For example, the interaction of two billiard balls on a billiard table can be treated as a constrained two-body problem, without containing any additional variables that could represent in principle possible additional influences on the motion of the two balls, due to other bodies or fields. That is, we represent the interaction between the two billiard balls as if it were taking place in a simple universe that contained nothing but the two billiard balls and the table and a gravitational field constraining the balls to the surface of the table.

Third, models may represent a phenomenon not in terms of a full-fledged initial value problem. For example scattering phenomena are often modeled as interactions between a scattered particle and a scatterer (and sometimes as an interaction between a microscopic particle and a macroscopic medium) without specifying a complete initial value problem, which would require a specification of the initial fields and of the microscopic states of any other charged particles in the vicinity.

Fourth, it is again important to remember the distinction between models and the real world systems they are intended to represent: In some sense the viscous fluid model of the ground surrounding the accelerator is extremely coarse-grained and partial, since it leaves out many of the details characterizing the rock formations in which the accelerator was built. But nevertheless the model provides us with the specification of the values of certain quantities at each spacetime point. That is, a model can be *mathematically* complete in Field's sense in that the model has no 'holes' in which the value of the model's variables remain unspecified, without the model being *physically* complete and aiming to representing completely and precisely the physical system modeled.[4]

I have argued that many of the 'human-faced' features that Woodward and Field attribute to causal representations are characteristic of models in physics as well. But have I really shown that these are features of models in *fundamental* physics? One might object that my example merely shows that modeling in *applied* physics shares crucial features with causal modeling elsewhere. But my point here is not merely that there are certain aspects of the experiment, such as the treatment of ground motion, are modeled in an extremely coarse-grained way. Rather I want to emphasize that all the heterogenous models of the various components of the experiment model

[4] Below I discuss the question whether such a complete specification of values makes it difficult to apply an interventionist conception of causation to such models.

some aspect of the goings on in certain spacetime regions in the backward lightcone of putative detection events in the particle detectors. That is, the way our best current physics models the experiments intended to confirm the Standard Model is not through a complete model specifying the state of the world at each spacetime point in the backward lightcone of the particle detectors, as Field would have it, but through a heterogenous mix of partial models, drawing on the resources of different theories at different levels of 'fundamentality'. If this modeling practice is 'merely' an instance of applied physics, then it seems that every experimental test of a theory will turn out to be an instance of applied physics!

At this point the causal skeptic will probably want to draw attention to a distinction between two different kinds of model of physical theories: on the one hand there the models actually constructed with the help of a theory's fundamental equations, to be used to represent actual physical phenomena, such as models of the LHC. Cartwright calls such models "representational models". On the other hand there are models which are hypothesized to satisfy the laws exactly; the laws, that is, are strictly true of these models. Some of these models represent extremely simple 'possible worlds'. These are the kinds of model physics students study and investigate in problem sets. Cartwright calls these "theoretical models". But a theory's laws, it is said, also has models assumed to represent highly complex possible 'worlds' imagined (but only imagined!) to be specified in complete and microscopic detail, but—to the extent that these models are meant to approximate worlds like ours—much too complicated to be expressible by us. It is the latter models that according to the causal skeptic are inhospitable to causal notions.

If, however, we follow van Fraassen in holding that a representational model is a model of a phenomenon only if it is *used* by us as a model of that phenomenon, then the complete possible worlds postulated by the philosopher to satisfy a theory's fundamental equations cannot be representations at all, since they are too complicated to be grasped, let alone used by us. Yet we do not have to fully endorse van Fraassen's view to be able reject the inference from the alleged inhospitability of complete dynamically possible worlds to the absence of causal notions in fundamental physics. No matter what the correct account of the role of the two kinds of model in physical theorizing is, it seems to me that Field (and perhaps Woodward as well) has been seduced by a powerful and widely accepted image of the content of physical theories according to which what a fundamental theory like classical electrodynamics or quantum field theory 'really' tells us about the world is fully embodied in hypothesized immensely complex solutions to the fundamental equations that are postulated to capture real world phenomena in complete detail. To assume that the content of a physical theory is exclusively given by these complete structures,

which we do not and cannot actually have 'in our hands', rather than by the ways in which physicists *actually use* a theory to represent actual physical systems seems to me to get things exactly backward.

The practice of modeling in physics, I have argued, shares crucial similarities with causal modeling elsewhere, and we should reject premise 2.5* in the argument above. Premise 2.6, as I want to argue now, can be questioned as well. (Ney 2009) argues that a notion of cause as difference maker, perhaps along the lines of Woodward's interventionist account, falls prey to Field's argument and that the only way to avoid Field's conclusion is to introduce a separate notion of cause that she calls "physical causation." But we can grant Field and Woodward that the distinction between salient causal factors and background conditions plays an important role both in ordinary discourse and in the special sciences and yet resist the claim that it is *essential* to the notion of cause as difference-maker.

The complete putative models to which Field draws our attention do not allow us to draw a principled distinction between cause and background condition. Yet that in itself does not make causal notions inapplicable to such models, if we insist that this distinction is an important *pragmatic* component of our concept of cause, which reflects the close connection between the concepts of cause and of explanation. Not everything that in principle can make a difference to the occurrence of a particular event is salient in every context: in some context the striking of the match is the factor on which we wish to focus, but we can imagine other contexts in which the presence of oxygen becomes the focus of our attention. Yet in addition to its pragmatic dimensions we can also identify a non-pragmatic (or perhaps more carefully 'less-pragmatic), minimal core of an asymmetric concept of cause, which includes any and all factors that could in principle make a difference to the occurrence of an effect. Moreover, that the distinction between causes and background conditions is heavily pragmatic, does not mean that it is merely subjective or arbitrary. Rather there may be very good 'objective' reasons for excluding a certain factor from the set of possible causes of an event. And including the entire backward lightcone does not imply that we have to treat all events in the past equally.

As we have seen above, Field worries that once the entire backward light cone of an event e is included in the set of its causes, then we have to treat all macro-events in the past of e as causally on a par: Sam's praying comes out as a cause of the fire's going out just as Sara's aiming the water hose at the fire. Woodward discusses another example in this context, which I want to use here to illustrate and examine Field's claim. Suppose that the event e is the event of my headache disappearing (represented by the variable E); that 30 minutes prior to that event (at t) I took

aspirin (event *a* represented by variable *A*), that around the same time *t* my neighbor sneezed (event *s* represented by *S*) and that I wished my headache would disappear (*w* represented by *W*). While taking aspirin may be an effective strategy for making a headache go away, bringing about events *w* or *s* are not. Yet, Woodward points out, if we consider very fine-grained microscopic specifications (*E**, *A**, *W**, *S**) of the states of the fundamental particles realizing events *e*, *a*, *w*, and *s*, respectively, then it will turn out that *E** does not only depend on an exact specification of *A**, but also on a specification of *W** and *S**, and in fact of the entire cross section of the backward light cone at *t*. According to Woodward, this threatens to undermine the whole point of the notion of causation. "If we are forced," he says, "to the conclusion that [sneezing and wishing] are causes of *e* as well, the motivation (according to manipulationist accounts) for introducing the notion of causation in the first place—the intuitive connection between causation and manipulation and the contrast between causal and merely correlational relationships—appears lost." (Woodward 2008, 84-5). But is the antecedent true and are we indeed forced to the conclusion that sneezing and wishing are causes of the headache's disappearance, once we posit not only a macroscopic model but also a putatively complete micro-model of the situation? The answer, as I want to argue now, is 'no'. For reasons that Woodward himself discusses in his paper (Woodward 2007, 85-87), we can preserve both the intuition that my neighbors sneezing is not a cause of my headache's disappearing, even though the precise microstate of my neighbor during his sneezing might be a cause of what the precise microstate of my brain is as my headache vanishes.

There seem to be two related worries concerning a complete putatively causal micro-model. First, one might worry that it simply follows from the fact that the *fine-grained* variables are causally related—that is, that *S** is a cause of *E**—that the *coarse-grained* variables are causally related as well. Second, the argument might be that once we grant that *E** depends on *S**, we have to worry about extreme 'fillings' for the values of the variables *S** that would affect the occurrence of *E*. In both cases we are led to conclude that *S* is a cause of *E*, even though we do not think that we can manipulate *E* by means of manipulating *S*. Thus, if we want to retain the connection between causation and manipulation, we have to deny that there is fine-grained causal dependence: on the micro-level there is nomic dependence but no causal dependence.

In reply to the first worry, we can concede that the *particular precise* disappearing of the headache *E** depends on the occurrence of *W** and *S**, without accepting that neighbors' sneezes cause headaches to disappear. Arguably a fine-grained causal model will imply that interventions into the micro variables involved in the characterization of *S** will result in changes to the micro variables involved in the characterization of *E**, yet the causal claim Woodward worries about is not

this but the claim that the value of a *coarse-grained* variable representing sneezing is causally related to the value of a *coarse-grained* variable representing the disappearance of headaches. Presumably, however, changes in the values of the micro variables characterizing *S** that are associated with a change in the macro 'sneezing-variable' *S* from one to zero will *not* result in changes in the values of the micro variables characterizing *E** that are associated with a change in the macro-variable *E*. That is, while the *microscopically precise* way in which the neighbor sneezes will make a difference to the *microscopically precise* way in which my headache disappears, whether my neighbor sneezes or simply sits there quietly will not make a difference to whether my headache disappears or not. Thus, both on the micro level and on the coarse-grained macro level the connection between causation and effective strategies for manipulation is preserved: The fine-grained causal model predicts that the precise way in which the sneezing occurs affects the precise way in which my headache disappears, and intervening into the precise way in which the sneeze occurs is an effective strategy for manipulating the precise manner of the disappearance of my headache. Yet since changes to how and even whether the sneeze occurs only affect the fine-grained details as to *how* my headache disappear, but not *that* it disappears, there is no causal relationship between the coarse-grained two-valued variables representing the occurrence of a sneeze and the disappearance of a headache. By the same token, intervening into whether neighbors sneeze is not an effective strategy for making headaches go away.

What about the second argument? The worry is that our discussion so far has ignored the possibility of there being extreme changes to the variables *S** which will result in my headache not disappearing. For example, we might assume that if my neighbor had emitted a sufficiently loud and shrieking noise instead of having sneezed, then my headache would not have disappeared. Nevertheless, we do not think that manipulating whether the neighbor sneezes can be part of an effective strategy for combating headaches, or so one might say.

Yet if we were to consider my neighbor's emitting a shrieking noise as a genuine possibility, then it seems that we ought to include in our coarse-grained causal model a variable that takes on different values whether the neighbor shrieks or does something more quiet such as sneezing. Whether my neighbor sneezes rather than shrieks will make a difference to whether my headache disappears. Thus, the neighbor's sneeze would indeed play a causal role in my recovery from the headache and, once again, the relation between causal claims and claims about interventions is preserved. If we are considering that changes to the neighbor's sneezing can affect whether or not my headache disappears, then intervening on the neighbor's sneeze in the right way will be a way of influencing the disappearance of my headache.

In many contexts, however, we can have very good reasons for excluding 'extreme fillings'—since we might know that my neighbor is not inclined to emit loud shrieks, or that the walls of his house are thick enough to block any but the very loudest shrieks—and in such contexts it is reasonable to adopt causal models that allow only for fine-grained dependence of E^* on S^* but not for coarse-grained dependence of E on S. To some extent, of course, it is a pragmatic question what possible values for the fine-grained or coarse-grained variables we are willing to consider in a given causal model. Yet this pragmatic dimension is part of the construction of causal models more generally: which particular causal model we choose in order to represent a certain situation is to some extent due to pragmatic and context-dependent factors. Moreover, as we have seen above, similar context-dependent constraints also guide our construction of dynamical models of a given situation.

Thus, Field's argument that putative complete microphysical models of a phenomenon could not be interpreted causally, along the lines of a notion of causation as difference maker, fails. To sum up my discussion of the argument in terms of the example cited by Field himself: Even though the *precise micro-state* of Sam's body comes out as a cause of the *precise micro-history* of the fire—as well it should—his praying is not a cause of the fire going out, as long as the relevant range of values of the variable representing his actions in our causal model includes only values representing his simply sitting there, performing a dance, or praying, and the like. Yet once we include a value for the variable constituting an 'extreme possible value', such as one representing his dousing the flames with gasoline, then his praying does indeed come out as causally relevant to the fire going out and plausibly so.

3. Interventions and Arrow-breaking

While Woodward couches his discussion of the problem of fine-grained dependence in terms of a manipulability conception of causation, such a conception is not essential to the arguments considered so far. In this section and the next I want to examine a range of arguments that focus directly on the notion of intervention, arguing that causal notions are inapplicable at the level of fundamental physics precisely because the notion of an intervention becomes inapplicable at that level.

Woodward's believes that various causal relations can be captured in terms of the responses of a system to possible interventions into the system. He defines the notion of *total cause* in terms of the notion of intervention as follows:

> (C) *X* is a total cause of *Y* if and only if under an intervention that changes the value of *X* (with no other intervention occurring) there is an associated change in the value of *Y*. **(Woodward 2007, 73)**

The notion of intervention, for Woodward, is itself a causal notion and is defined as follows:

> *I* is an intervention variable on *X* with respect to *Y,* if and only if, *I* meet the following conditions:
>
> (IV)
>
> 1) *I* causes *X*.
> 2) *I* acts as a switch for all the other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend upon the value of other variables that cause *X* and instead only depends on the value taken by *I*.
> 3) Any direct path from *I* to *Y* goes through *X*. That is, *I* does not directly cause *Y* and is not a cause of any causes of *Y* that are distinct from *X* except, of course, for those causes of *Y*, if any, that are built into the *I*-*X*-*Y* connection itself; that is, except for (a) any causes of *Y* that are effects of *X* (i.e., variables that are causally between *X* and *Y*) and (b) any causes of *y* that are between *I* and *X* and have no effect on *Y* independently of *X*.
> 4) *I* is independent of any variable *Z* that causes *Y* and is on a directed path from *I* to *Y* that does not go through *X*. **(Woodward 2007, 75)**

The guiding idea is that by intervening into a system we can 'wiggle' the value of a variable independently of what other causes it might have and that tracing the effects of such wiggles through the system in question allows us to determine the causal structure of the system. Woodward's account obviously cannot provide a conceptual analysis of the notion of total cause in terms of that of intervention. Since the definition of an invention does itself appeal to causal notions, the analysis would be circular. Rather, Woodward's explicit aim is to elucidate the concept of cause by making precise various conceptual connections between that concept and several related concepts, such as that of intervention or of counterfactual dependence.

For an interventionist account of what it is for *X* to cause *Y* to be applicable in physics, the counterfactual describing what would happen to a variable *Y* within a physical model under an intervention on *X* has to have a determinate truth value (Woodward 2007, 91). But, Woodward worries, there are various reasons for doubting that these counterfactuals are well defined, and,

hence, that the notion of intervention can be meaningfully applied to models constructed with the help of fundamental physical theories.

Woodward's first worry is that interventions have to be 'arrow-breaking' but that models of fundamental physical theories do not allow for arrow-breaking changes to a model. Woodward's motivates the condition of arrow-breaking by pointing out that causal relations in the special sciences only have a limited range of invariance. At least some interventions thus can be thought to correspond to a change to the system that takes it outside of the realm in which the causal relations originally postulated hold. Thus, such an intervention interrupts any causal dependence of the variable intervened upon on any other variables aside from *I*: interventions break any causal arrows ending in the intervention variable aside from that from the intervention variable itself. If, by contrast, we take the range of invariance of the dynamical relations governing a system to be unlimited, then I might appear doubtful whether there could be any arrow-breaking interventions. While interventions might *change* the value of *X* from what it would have been otherwise, an intervention would not be able to interrupt *X*'s dependence on its other causes and, hence, what that counterfactual value *X* would have, would still depend on the values of *X*'s other causes.

But one may ask whether all interventions that allow us to probe the causal structure of a system have to arrow-breaking and whether it might not be enough that an intervention change the value of the variable intervened upon. This will be one of the replies to Woodward's worry I will develop in what follows. There are two other replies: first I will argue that Woodward's switch condition, condition (IV.2), can be satisfied even for interventions that are not arrow-breaking. Thus, even if it were true that representations in physics do not allow for arrow-breaking interventions, this does not mean that Woodward's notion of an intervention would not be applicable. Second, I will argue that an arrow-breaking notion of intervention can be legitimately applied to at least some models in physics. And third, I will argue that the applicability of causal notions does not depend on a notion of intervention that satisfies the switch-condition and that intervention variable that merely change the value of the variable intervened upon without breaking all other causal arrows also allow us to probe for causal structures.

According to condition 2 in Woodward's definition, an intervention acts as a 'switch': an intervention on a variable *X* renders *X* independent of all other variables that cause *X*. Consider, first, as a simple example a model of the trajectories of a (small) number of billiard balls on a pool table. Let the variable *X* be the position of ball *x* at time *t*, which according to the model depends on *x*'s prior trajectory and any collisions with other balls. Let the putative intervention consist in my giving the ball an extra nudge with the cue. My action will change the position of ball *x* from what it

would have been without my nudging it, but it does not render the ball's position independent of its other causes, such as the earlier motion of other pool balls: the final position of the ball depends *both* on the nudge it receives from the cue *and* on the ball's prior state of motion. Thus this 'intervention' does not satisfy Woodward's switch condition. Yet since the example presents what intuitively seems to be an intervention or manipulation, one might ask if Woodward's notion of an intervention is not too restrictive and whether it is possible to formulate an interventionist account of causation in terms of a broader notion of intervention, which can cover cases such as this example as well. I will return to this question below. The first point to which I want to draw attention is a different one: while every arrow-breaking intervention acts as a switch, the converse does not hold and there can be switches that satisfying condition 2, which are not arrow-breaking.

Imagine as a slight alteration of the example that I reach into the table to place the ball *x* at some specific location determined by me, independently of its present state of motion. My action now does act as a switch in Woodward's sense, in that it renders the position of the pool ball counterfactually independent of it prior state of motion and therefore constitutes a genuine intervention, according to (IV). We could represent this intervention in a causal model in two different ways. First, we could represent it as a genuinely 'arrow-breaking' intervention of the kind Woodward has in mind—that is by a model in which the only causal arrow into *x*'s position is from the intervention variable and there are no causal arrow into *x*'s position from the past of *x*'s trajectory. **[see figure 3]**. But alternatively we could take the values of the intervention variable representing my action to depend on the prior state of the system. The intervention is set up in such a manner that it imparts just the right motion on ball *x* that, no matter what its prior motion, it ends up at the specified location. But in order to achieve that result, my action has to take the prior state of motion of the ball into account: I have to reach into the table at the right location and change the momentum of the ball in just the right way. That is, instead of 'breaking' the causal arrows going into *X* an intervention acting as a switch could be represented as a feedback system that involves additional causal arrows from the causes of *X* into *I*, which 'measure' the values of *X*'s causes to ensure that there is no counterfactual dependence of *X* on any of these causes. This kind of intervention is best represented as involving two intervention variables: a control variable I_C that controls the value to which we want to set *X* and that satisfies condition (IV) and a second feedback variable I_F that represents the feedback mechanism and whose values also depend on the values of *X*'s causes. **[figure 4]**

Thus, even though Woodward himself thinks of interventions as arrow-breaking, there is another way of satisfying the switch condition (2), by modeling an intervention through the

addition of a feedback-system that renders the value of *X* counterfactually independent of its other causes, even though no causal arrows are 'broken'. Hence, a first response to Woodward's worry that the "idea of breaking some causal influences to investigate others" (94) does not carry over to fundamental physics is this: there can be interventions that act as switches but do not break any causal arrows.

Modeling an intervention as adding a feedback system might appear to be more realistic from the perspective of the underlying physics in that it does not have to abstract away from the dynamical dependence of the motion of the ball on its prior state in our example, yet the model is not entirely unproblematic. A causal model in which the causal dependencies along different routes exactly cancel—and hence that is such that the causal dependence of *X* on (some of) its causes is not reflected in a probabilistic dependence of *X* on the causes in question—does not satisfy a condition that is frequently imposed in the causal modeling literature, that any cause of a variable *X* is probabilistically correlated with *X*—or, in the terminology introduced by Patrick Suppes **(Suppes 1970)**, every cause of *X* is a *prima facie cause* of *X*. Judea Pearl calls such an assumption *stability* (see Pearl , 48), while Spirtes, Glymour and Shines call it *faithfulness* **(Spirtes, et al., 1993)**. But it is unclear whether the correct conclusion to draw from this is that one cannot model switches as feedback systems (since this would violate stability), or whether instead we should conclude that the stability condition cannot be a general constraint on all causal models.

(**Cartwright 2001**) argues that we cannot expect systems violating stability to be uncommon, as does Kevin Hoover **(Hoover)**, who argues that violations of stability occur frequently in macroeconomics. Indeed, there seem to be many examples of feedback systems, both naturally occurring and constructed by humans, for which a representation in terms of a causal structure that violates stability would be a natural choice. As an example from biology, consider body temperature, which (within a certain range of invariance and to a certain degree of approximation) is counterfactually independent of ambient temperature, precisely because the human body responds to changes in ambient temperature through different mechanisms along different causal routes, whose function is to maintain a constant temperature. The external heat transfer mechanisms are radiation, conduction, convection and evaporation of perspiration. These mechanisms do not merely operate passively, but are exploited by by neural feedback mechanisms. For example, sweating begins almost precisely at 37°C and increases as the skin temperature rises above this value; and there are a variety of mechanisms directed at conserving body temperature, such as shivering to increase heat production in the muscles or vasoconstriction to decrease the flow of heat to the skin.

The LHC once again can provide us with an example of a sophisticated feedback system that is best represented in terms of a stability-violating feedback-model of interventions **(reference to Steinhagen)**. There are various sources of disturbance of the proton beam, both environmental and accelerator-inherent. These disturbances can lead to a loss of particles inside the machine, which has the potential of severely damaging the accelerator. Thus it is important to counteract these disturbances, which is done through feedback systems designed to keep the orbit stable by controlling key beam parameters, such as its orbit, its energy, and its tune (defined as the number of transverse oscillations a particle describes per turn). The size of the LHC feedback system is quite impressive: the beam control system consists of over 1000 beam position monitors and over 1000 orbit corrector magnets, whose fields are adjusted according to calculations of a central global feedback controller. This control system is naturally represented as causally intervening into the proton beam to keep it stable and to prevent damages to the machine: The beam is one physical system with a certain time-evolution into which a second system—the beam control system—intervenes. Indeed, **(Steinhagen)** explicitly represents the control system causally in terms of box-and-arrow-diagram in which "the signal flow and causality are indicated through arrows." (40) **[figure 5]** Even though the effect of the control system (ideally) is to render the variables characterizing the orbit independent of the size of prior perturbations, as Woodward's condition (2) on intervention requires, the intervention does not proceed by breaking the 'causal arrows' going into the beam variables but by being extremely sensitive in its responses to measurements of the prior state of the system intervened upon.

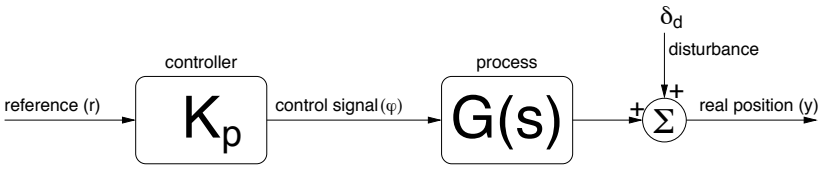

Figure 5

Now one might argue in defense of the stability condition and as a criticism of my feedback model of interventions that we can never realistically expect exact cancellation along different routes.[5] Body temperature will not stay *exactly* constant as ambient temperature changes and the controlled beam parameters in the LHC are not *completely* independent of the prior particle orbits.

[5] See also the discussion in (Cartwright 2001, 253-4).

Thus, one might want to insist that the stability (or faithfulness) condition can be saved and that feedback systems of the kind I posited do not offer a genuine alternative model of interventions to an arrow-breaking model, since they do not in fact satisfy the 'switch'-condition. Again, however, it is important to distinguish carefully between causal models and the real world systems they are intended to represent. If a feedback control system works so well that the quantity controlled remains constant within the bounds of experimental error, or if the variations in the quantity are small enough to be ignored given a certain context and scientific interest, then it seems to be appropriate to represent the quantity at issue in our causal model by a variable whose values are strictly independent of the variable's other causes, given the operation of the feedback system. If we nevertheless were to insist that the causal model has to satisfy stability, then this would require us to introduce a more fine-grained dependence of the variable on its causes than is warranted either by our measurement precision or by the context in which the model is used.

Moreover, appealing to fine-grained dependencies in order to rescue stability leads to an odd dialectical situation as far as our main issue is concerned—the place of causal notions in physics. For the objection seems to imply that the *stability condition* can be expected to hold in general only once we move to especially fine-grained models of physics and not for the much more coarse-grained and arguably more highly idealized models of the higher sciences. To the extent, then, that stability is a necessary condition for the applicability of causal models, fundamental physics seems to be much more hospitable to causality than the higher sciences, where, as our models get ever more coarse-grained and idealized, we would expect failures of stability to become more and more common. By contrast, Woodward's *switch condition* seems to be satisfied only in the higher sciences, if the objection is to be believed.

I have argued that Woodward's switch condition can be satisfied even in models of interventions that do not involve arrow-breaking. But, and this is my second response to Woodward's worry, it is far from clear that models that do involve arrow-breaking are not applicable in physics. To reiterate a point I have by now made repeatedly, a causal model of a physical system need not be complete and represent all quantities and interactions that are part of the system, if these do not play a significant role in the aspect of the system's behavior in which we are interested—any model we construct, no matter at which level of 'grain' will abstract and idealize in some way or other. Woodward maintains that "in many physics contexts there may be no physically realistic operation corresponding to placing some variable of interest entirely under the control of an intervention variable and breaking all other causal arrows directed into it." (**Reference?)** But if we keep in mind that what is at issue is whether there are "physically realistic"

circumstances which*, in a given context, are appropriately represented* as involving an intervention variable that breaks all other causal arrows directed into the variable at issue, then it is not so difficult to imagine such circumstances. For example, an arrow-breaking model may be appropriate in cases in which the intervention under consideration effectively swamps all other influences on the variable intervened upon and therefore allows us to ignore any in-principle dependence of the variable's values on its other causes. Again we can illustrate this with an example from the LHC. Certain failures in the LHC can trigger a 'beam dump'—that is the extraction of the proton beams from the accelerator into a large absorber. The triggering mechanism arguably ought to be represented as an intervention that breaks the causal arrows from the goings on in the accelerator to the states of the different particle detectors. Even if we were to grant Field's point that events in any of the detectors at one time *in principle* depend on everything in the events' backward lightcones and, hence, also on the state of the accelerator at earlier times even when no proton beams are present in the accelerator, we do not need to explicitly represent this in-principle-dependence in our models of the accelerator and detectors. Rather, we can assume that activating the beam dump effectively breaks any causal arrow from the accelerator to the particle detectors.

So far I have taken Woodward's notion of intervention for granted but have argued that the notion of an intervention variable acting as 'switch' is applicable to systems of physics. Yet even if condition (2) was not applicable to many systems of physics, this does not entail that we could not represent these systems causally or would have to give up on the close connection between causal notions and the notion of manipulation or intervention. For, and this is my third response to Woodward's worry, there is an alternative strategy available and that is to argue that Woodward's notion of intervention is too narrow. Retaining Woodward's idea that interventions are associated with an intervention variable, we might replace condition (2) with the weaker condition

2*) Certain values of I are such that when I attains those values, the value of X is different from what it would have been without the intervention.

That is, all we require of an intervention is that it be able to *change* the value of X from what it would have been otherwise and not that it do so by breaking all other causal arrows leading into X. This still allows us to use possible interventions as a guide to causal structure: we can ask if an intervention that changes the value of X and satisfies condition 3) and 4) results in a change in the value of Y. Again the LHC can provide an example. Slow particle losses in the machine are minimized with the help of a series of collimators that absorb these particles. Depending on the precise trajectory of a slow particle, the primary collimator will either absorb the particle or scatter it so that the particle is absorbed by a secondary (or even tertiary) collimator. That is, the beam

cleaning system intervenes on the slow particles of the beam, but in ways that depend on the state of the incoming particles: the trajectory of a particle depends both on its prior state of motion and on the action of the collimator. Despite the fact that this intervention does not act as a switch, it has the marks of a paradigmatic intervention and the particles' trajectory is counterfactually dependent on the action of collimator.

It is unclear why we would want to insist that interventions have to be arrow-breaking or at least would have to satisfy the switch condition. An advantage of arrow-breaking interventions is that they can be represented mathematically very simply: if we represent causal relationships through structural equations, then we can replace the equation for the variable intervened upon simply by one that assigns that variable a specific value. But the insight that there is a close connection between causation and the notions of intervention does not seem to require arrow-breaking interventions. All that does seem to be required to distinguish mere correlations from causal relations is that an intervention on a variable *X* can make a difference to the values of variables that are effects of *X* but makes no difference to the values of variables correlated with *X* without being effects of *X*.

Moreover, there is an additional advantage of allowing for non-arrow-breaking interventions, for there is a general worry about wedding the concept of causation too closely to that of arrow-breaking. An arrow-breaking intervention can only be a reliable guide to the causal structure of the system prior to the intervention if we assume that causal structures are modular—that is, that all the remaining arrows, and in particular all those from *X* to *Y* remain unaffected by the arrow-breaking. But while plausibly many causal structures are modular, it is far from clear that modularity is satisfied in general **(reference to Cartwright)**. A non-arrow breaking intervention, by contrast, merely hitches a new intervention variable to an existing causal structure that is otherwise left intact and demanding that such addition to the causal structure are possible is weaker than the condition of modularity.

To sum up, in this section I have responded to Woodward's worry that the condition that interventions are arrow-breaking will not be satisfied in models constructed with the help of fundamental physical laws. My response was three-fold. First, I argued that there can be non-arrow breaking interventions that nevertheless act as switches and render the variable intervened upon counterfactually independent of its other causes. Second, I argued that once we grant that all scientific representations idealize and distort in some ways or others, there is no in principle obstacle to representing physical interactions in terms of genuinely arrow-breaking interventions, even in cases where we believe that an in principle dependence on the severed causal links persists.

What is more, there arguably will be many cases in which the most appropriate representation in a given context will be one that ignores the in principle dependence of certain variables on other influences that have been effectively shielded off. Finally, I argued that there are reasons to believe that Woodward's notion of an intervention that acts as a switch is too narrow and that a notion of intervention that merely postulates changes to the value of a variable intervened upon without breaking all other causal links has certain advantages over Woodward's notion.

4. Large and small worlds

I now want to turn to a cluster of worries concerning the applicability of the notion of intervention that are associated with putative spatio-temporal properties of the representations which we construct with the help of our physical theories. First, it seems to follow directly from Field's claim that physics provides us with models of a spatial region R at time t that require as input an entire cross section C of the backward lightcone of R that the notion of intervention cannot be applied to models in physics. For if physical models consist in entire backward lightcones (up to the cross section C on which the initial data are specified), then interventions into the model are physically impossible. **[see figure 5]** Any putative intervention I_{out} originating outside of the backward lightcone could not reach R at t without involving faster than light propagation, while any variable I_{in} representing an event inside of the backward lightcone of R would already have to be included in the model as one of R's putative causes, and therefore could not represent an external intervention into the system. Thus, interventions into backward-lightcone-models are physically impossible.

In light of our preceding discussion my reply to this worry will be obvious: many, or even most, models constructed even with the help of our most fundamental theories do not involve specification of the state of the full backward lightcone of the spacetime region we are modeling, but also include spatial boundary conditions. Thus, even if we accept Field's and Woodward's claim that the representation of the system does not involve any spatial or temporal gaps, this does not imply that there is no room for interventions into physical systems: interventions can occur across the spatial boundaries of the system, and can be represented as a localized change in the boundary conditions.

My reply presupposes, however, that the system in question has an outside from which we can intervene in the system and the legitimacy of this presupposition has been questioned by both Woodward and Hitchcock. At least some of our mature physical theories also provide us with models representing the universe as a whole and such models, it seems, leave no room for an external vantage point or an 'outside' from which to intervene into the model.

That it is difficult or even impossible to locate causal relations within theories that have potentially universal scope, is a view shared by many writers who have sympathies for interventionist accounts of causation. Thus Pearl maintains that "if you wish to include the whole universe in the model, causality disappears because interventions disappear" (2000, 350); Christopher Hitchcock claims that "how to find or even understand causation from within the framework of a *universal* theory is one of the very deep problems of philosophy" (Hitchcock 2007, 53); and even though Woodward is more tentative in his skepticism, he also believes that causal ascriptions become "increasingly strained when candidate causes expand to include the state of the entire universe" (Woodward 2007, 93).

The worry, as both Hitchcock and Woodward state explicitly, is both epistemological and conceptual: what reasons can we have for attributing causal relations to a model of the universe as a whole systems? And how do we understand putatively causal relations that cannot, even in principle, be exploited through an external intervention?

I want to address the second worry first. Alexander Reutlinger **(reference)** argues that Woodward's definition of total cause does not require that an intervention be physically possible but merely that it be logically possible. If we allow that it would be logically or conceptually possible for the universe to contain more objects or quantities than it actually contains, then a counterfactual intervention into a variable that is part of a global model (representing the entire content of the actual universe) might be modeled in terms of an additional variable representing some object that we posit counterfactually to exist in addition to the universe's actual content. While Reutlinger's point seems to be correct in principle, his strategy of counterfactually adding an additional intervention variable to an existing global model seems not be be available in the case of global models defined over all of spacetime. In this case there is simply no location available from which the invervention could occur.

There is a more promising strategy available, however, and that is to invoke an alternative notion of intervention that does not require positing an additional intervention variable outside the system intervened upon and, hence, can be unproblematically applied even to a global model of the universe as a whole. Ironically, Pearl himself formally introduces such a notion in (Pearl 2000, 69f.). A *causal model*, according to Pearl, consists of a set of variables {*X*, *Y*, ...}, a directed acyclic graph, and structural equations that specify the value of each variable in terms of the value of the variable's causal parents. An *intervention* on the variable *X*, according to Pearl, then consists in removing the structural equation for *X* from the causal model and replacing it by an equation that fixes the value of *X* to some non-actual value *x*. Pearl calls such an intervention an "atomic

intervention", denoted by *do*(*x*) (Pearl 2000, 70). While we might think of the change in the structural equation as the result of coupling *X* to some external intervention variable, this is not part of Pearl's formal account. Thus, contrary to Pearl's own worry, it makes perfect sense to ask, on his account, how the values of variables would change under an intervention into a model of the universe as a whole.

One might worry, however, that the resulting counterfactuals require that we investigate nomically impossible situations or worlds. The dynamical laws of many of our mature theories are both past- and future-deterministic. That is, for any two dynamically possible worlds W_1 and W_2, W_1 and W_2 agree in their state at some time *t*, exactly if $W_1 = W_2$. Thus, if the value of some variable changed as a result of an 'intervention', then dynamical laws entail not only that the evolution of the system in the future of *t* will be different from the system's actual evolution, but that its past evolution would have to have been different as well. Conversely, if we insist that interventions can only have an effect on the (causal) future evolution of a model, then one might worry that the dynamical laws cannot be applied in a meaningful or coherent way to the model, since the model is then posited to have a past that is dynamically inconsistent with its state after the intervention, and that, hence, counterfactuals concerning the effects of intervention cannot have well-defined truth-conditions. In the case of finite systems, as Woodward points out, modeling interventions with the help of intervention variables allows us to keep nomic violations 'off stage' and restrict them to occur outside of the system of interest. In the case of global models this move is not available. An intervention results in a dynamically impossible model and this, one might worry, threatens to render the whole approach incoherent and not-well-defined.

But once we posit a *causal* model for the universe, the model provides us with a well-defined procedure as to how to apply the dynamical laws in evaluating the results of interventions, even if they involve a violation of the dynamical equations governing the system. Pearl's notion of an intervention, like Woodward's, involves arrow-breaking in that the structural equation for the variable *X* intervened on will be replaced by one that assigns *X* some non-actual value. Since the arrow-breaking concerns only the *causal past* of *X,* we can use the original model, including information about the dynamical laws encoded in the structural equations, to determine what the *causal future* of the system would be from the intervention on *X* onward. In other words, we can use the state of the world at the time of the intervention as input in the dynamical laws to determine the future evolution of the world. But because the intervention changes the causal structure *leading into X,* we cannot apply the dynamical laws to evolve the state at the time of the intervention backward in time. That is, because the intervention is nomically impossible we cannot

apply the dynamical laws *across* the posited intervention, but this does not make the laws inapplicable to the model in general.

Intervention-counterfactual, therefore, are time-asymmetric, just as Lewisian counterfactuals are. But, instead of attempting to ground the causal or intervention asymmetry in a counterfactual asymmetry, I am proposing that we turn Lewis's account on its head and ground the asymmetry of intervention-counterfactuals in the asymmetry of the causal model within which they are evaluated. **[I will return to this issue in chapter ???]** Moreover, contrary to Lewis's view, I do not believe that there is a *standard* kind of context for evaluating counterfactuals in which backtracking counterfactuals are false. Rather, we can distinguish at least two perfectly standard contexts that play a role in the sciences: a context in which counterfactuals are evaluated asymmetrically as causal intervention-counterfactuals and one in which counterfactuals are evaluated purely dynamically by taking the state of a system or the world at one time as input and determining both the past and future evolution of that state in accord with the dynamical laws. If the laws in question are time-reversal invariant, then the counterfactuals will be symmetric and changes at one time will be associated with changes in both the past and future evolution of the system.[6]

I have argued that *once we posit a causal structure*, then intervention counterfactuals can have well-defined truth-conditions even in global causal models. I have not yet addressed the epistemological worry, however: can we ever have good reasons for postulating such causal structures within the context of a theory with universal scope? Yet this worry dissolves once we realize that even theories with potentially universal scope can be—and are, as a matter of fact—used to model finite systems with an 'outside.' For models of finite systems we can apply either Pearl's notion intervention discussed above or invoke Woodward's notion of intervention, which Pearl also later in his book characterizes as a "more appealing" account, and treat interventions in terms of exogenous variables. In either case, the kinds of reasons we can have for positing a particular causal model for a system will be analogous to the reasons we give for systems modeled by strictly domain restricted theories. That is, in answer to Hitchcock's worry, we "find […] causation from within the context of a universal theory" in the very same in which we find causation in theories with a limited domain: through interventions into finite systems modeled with the help of that theory.

---

[6] There may be yet a third kind of context in which the 'statistical mechanical' counterfactual introduced by David Albert and Barry Loewer is the appropriate one to use. I will discuss Albert's and Loewer's views in chapter ??.

If we assume that there are cases where we are justified in interpreting such *local* models causally, then a skeptic about causal relations in *global* models must believe that causal relations invariably disappear when a local model is embedded into a larger model of the universe as a whole. But it is not exactly clear what reasons might be offered in support of this belief. We can imagine the process of embedding local, finite models into successively larger models. Now, one might hold that at some point this process of enlarging reaches the limit of what can successfully be modeled even with a theory with putatively universal scope. This is Nancy Cartwright's view, who would caution that the fact that we can formally and abstractly speak of the possible worlds allowed by a theory should not mislead us into thinking that we are justified in accepting a theory as providing us with anything but local representations of particular phenomena. Cartwright argues that we do not know whether physics satisfies the principle of self-closure, that is, that there are "(in God's great Book of Nature) laws of physics that dictate everything that happens that can be reasonably taken to be in the domain of physics itself." (reference: Natural Laws and the Closure of Physics) But let us assume, for the purposes of the present argument, that we can indeed embed a given theory's local models into ever larger models and eventually into possible worlds.

The first question, then, is whether a causal model C associated with a local dynamical model D can be recovered as causal submodel from a larger causal model C^* associated with the larger dynamical model D^* into which D is embedded. Take a given dynamical model D with specific initial and boundary conditions. We obtain an embedding model D^* defined over a larger spacetime region if we treat the initial and boundary conditions of D themselves as dynamically determined in the larger model. If finite dynamical models of a theory suggest a particular causal interpretation, then the embedding requirement says that this interpretation can be preserved when we embed the model into a larger model. As long as we do not allow values for the variables newly added to the expanded model C^* that disrupt the causal relations within the submodel C, the causal relations comprising C will be preserved in C^*. As an example, think of a causal model of a car engine. The causal relations characterizing this model will be preserved when we embed it into a causal model of the car as a whole. Clearly there will be some settings for the variables external to the submodel representing the engine that preserve the causal relations within the submodel—among those will be the settings representing the actual situations in which we are warranted to adopt the causal model of the engine's functioning.

Now, the causal model for the engine presumably will only be invariant under some limited range of values of the variables characterizing its environment. Thus one might worry that all I have said so far is that causal models can be recovered as submodels from larger models into which

they are embedded exactly when they are thus recoverable. But local causal models of putatively universal theories are distinguished precisely by the fact that they have a very extensive and perhaps even unlimited range of invariance. In particular, if, with Field we were to posit relativistic models in which the state of an entire cross section *H* of the past light cone of an event *e* is taken to be the cause of the event in a *local* model *M*, then embedding the model defined over the section of the past light cone of *e* that lies to the future of *H* into a model *M** defined over a larger region of spacetime cannot affect the causal structure associated with *M*.

The only remaining question is what happens at the end point of our procedure of successive embeddings. Pearl, Woodward, and Hitchcock and others seem to suggest that it is at this very last step, when we embed a model of almost the entire universe except for a very small region in the models causal past into a model of the universe as a whole that all causal relations disappear. But if, as the current argument grants, we can justifiably interpret each finite sub-model of the model causally, and if the causal interpretation survives through an arbitrary number of stages of successively embedding a model into larger models, it is difficult to see what the argument for the claim could be that the global model could not similarly be interpreted causally.

The embedding condition I have just described is related to Woodward's modularity assumption: If the causal structure associated with the global model is modular, its substructures will survive a breaking-up of the global model into smaller local models, which are associated exogenous intervention variables, and, hence, reasons for attributing certain causal structures to local models will *ipso facto* be reasons for attributing a causal structure to the global model. But the modularity assumption has a different status in my argument than in Woodward's account: spelling out the notion of total cause in terms of arrow-breaking interventions requires that the modularity assumption holds for *all* causal (sub-)systems. By contrast, in order to refute the claim that we can never find causation in global models of our putatively universal theories, I only need to make it reasonable that the global causal models of *some* of our universal theories are modular. Moreover, Woodward's account requires 'modularity all the way down'—that is, that we can always in principle perform an arrow-breaking intervention on any variable in the model without perturbing the rest of the causal structure. My argument only requires that a global causal model is modular up to a point: it must be possible to break up the global model into local causal models in a way that satisfies modularity, but it may be that this process may reach a limit beyond with the local models cannot be broken up further without changes in the causal structure within the resulting submodels.

I have argued that considering interventions into local models of a universal theory can underwrite a causal understanding even of the global models of such a theory. Empirical interventions into a system offer not the only reason for attributing a causal structure to a system. As I will argue in detail in **(chapter 3?4???)** the existence of certain kinds of correlations between spatially separated events and asymmetries between prevailing initial and final conditions can also justify causal assumptions.

5. Field theories, Green's Functions and the Asymmetry of State Preparation

There is one final worry concerning the applicability of the notion of intervention in physics that I want to discuss. In my earlier discussion I focused on the switch condition (IV.2). Woodward also argues that condition (IV.3) is problematic in phyiscs, which states, intuitively, that for an intervention on *X* with respect to *Y, Y* may not depend on *I* along any causal route not going through *X*. At the level of a complete micro-model of a system the very notion of a causal path may break down and we may not be able to distinguish and isolate from one another dependencies along different routes. At the level of fundamental physics, Woodward suggests, the values of variables affect each other in such interconnected and interwoven ways that the only possible causal representation, if there is one, would be a representation according to which the complete state of the universe on one initial value surface causes the complete state at later times.

This worry, it seems, is particularly acute within the context of field theories or of theories of continuous media. The challenge is how to make sense of the idea of an interventionist notion of causation within the context of such theories: if we could no longer distinguish different causal routes connecting variables representing events on time slice to variables representing events on another and were forced to take the relata of the causal relation to be complete time slices in a model, then one might worry that the concept of causation would have become so diluted to have lost its usefulness. I want to propose two possible responses to this challenge. First, even if the causal relations relates entire time slices, an interventionist conception still allows us to capture one core aspect of the causal relation: its asymmetry; and time-asymmetric causal representations can play a useful role in physics even in cases where we do not represent a system as having a fine-grained causal structure. Second, I will propose a framework for capturing causal dependencies more fine-grained than entire time-slices even within the context of field theories.

Formal interventions into a causal model exhibit an asymmetry: they lead to changes in the values of variables causally downstream from the variable intervened upon, but not to changes causally upstream from the intervention. Just such an asymmetry also characterizes our

experimental interactions with physical systems, if we assume that the causal asymmetry lines up with the temporal asymmetry. Experimental interventions into an otherwise closed physical system affect the future evolution of that system but not its past and this is so, even when the dynamical laws governing the system are time-reversal invariant. Since the asymmetry in question is one associated with possible (and actual) interventions into a system, it is best thought of as a causal asymmetry. In explicit premise-conclusion form, my argument is this:

5.1 There is a temporal asymmetry characterizing experimental interventions into otherwise closed systems.

5.2 If there is such an asymmetry, it is best represented as a causal asymmetry.

5.3 Therefore, asymmetric causal representations play a legitimate role in physics.

Why should we accept 5.1? It is a striking fact about experimental interactions that we can only intervene into a system 'from the past,' as it were. Consider a system S that is governed by both past and future deterministic laws. That is, let us assume that the final state $S_i(t_f)$ of the system is uniquely determined by the initial state $S_i(t_i)$, where $t_i<t_f$, together with the dynamical laws and the boundary conditions; and that the initial state $S_i(t_i)$ is similarly determined by the final state $S_f(t_f)$. Thus, if S is closed between t_i and t_f, , then the initial and final states mutually determine each other. Nevertheless there is an asymmetry of state preparation in the following sense. We can prepare the system in its initial state $S_i(t_i)$ without making use of any knowledge we might have of the system's dynamical evolution between t_i and t_f; and we can subsequently calculate the system's future evolution for times $t>t_i$ from the initial state, the dynamical laws, and the boundary conditions. But we could not similarly first prepare the system's final state at t_f without using our knowledge of the dynamics and then take the final state together with the laws to calculate the system's past evolution for $t<t_f$. (Of course we cannot first prepare the system in S_f and then let it *evolve* into S_i. That is not what the asymmetry consists in. Rather the asymmetry consists in the fact that we cannot first prepare the system in S_f without *making use* of facts about the dynamical evolution and *then calculate* what the system's past evolution from S_i to S_f must have been, given the dynamical laws and the boundary conditions.)

There are two ways in which we can prepare the system in a specific final state at t_f. First, we can make use of our knowledge of the dynamical laws to determine the initial state in which the system has to start out at t_i in order to evolve into the final state in question and prepare the system in the appropriate initial state. In that case our ability to prepare the system in its final state relies crucially on our knowledge of the system's evolution between t_i and t_f. If the system is closed between t_i and t_f, we need to know which state S_i will evolve into the state S_f we are trying to set up,

because we can only 'prepare' the system in the state S_f by preparing it in S_i. By contrast, we can prepare the system in an initial state S_i without any knowledge of the dynamical evolution of the system between t_i an d t_f, even though the dynamical laws and boundary condition determine what that evolution is. While, given the dynamical laws and boundary conditions, the initial and final states determine each other, we do not need to make use of that fact if our goal is to set up the system in some specified *initial* state. But we do need to make explicit use of the dynamical evolution between t_i and t_f if our goal is to set up the system in a specified *final* state.

We can imagine, for example, that one experimenter is responsible for preparing a system S in an initial state S_i or a final state S_f and that a different experimenter is responsible for setting up the boundary conditions. If the first experimenter wants to prepare the system in a certain initial state, he can do that without knowing what boundary conditions the second experimenter chooses to set up. But if the first experimenter wants to make sure that the system ends up in a certain final state, he needs to know what the boundary conditions will be in order to make sure he prepares the system in the appropriate initial state.

A second way of preparing the system in a final state S_f is to prepare the system in that state directly by intervening into the system between t_i and t_f. In this case we do not need to make use of our knowledge of the evolution between initial and final times. But then the system will not be closed between t_i and t_f and we cannot use the dynamical laws and boundary conditions governing the closed system to retrodict the initial at t_i. Thus, there is a way for first experimenter to prepare the system in S_f without knowledge of the boundary conditions which the second experimenter tries to set up, but at the cost of having to violate these conditions through his intervention and, thus, by losing any ability to retrodict the evolution of the system with the help of the dynamical equations.

That is, even systems that are governed by both past and future deterministic dynamical equations exhibit an asymmetry of state preparation. In the case of a system that is closed between t_i and t_f and initial and final sates mutually determine each other, given the dynamical equations and boundary conditions, we can only prepare a system in a given final state S_f by making explicit use of the dynamics and prepare the system in the corresponding initial state S_i. If instead we directly intervene on the state of the system at t_f, we can no longer use the state at t_f to retrodict the state at t_i (while we can, of course, directly intervene on the state of the system at the earlier t_i and then predict the system's evolution until some later time t_f). We can only intervene into a system from its past.

The asymmetry of state preparation is a paradigm case of the causal asymmetry, as understood by interventionist accounts of causation. In particular, if S_f is an effect of S_i, then according to an interventionist account of causation there are two ways by which one can intervene on the system to set S_f to a particular value: first, we can intervene on S_i, which in turn will affect the value of S_f; or, second, we can intervene directly on S_f, perhaps by breaking the causal arrow from S_i to S_f and making it impossible to retrodict the value of S_i on the basis of the value of S_f. Thus, interventionist accounts of causation predict that experimental systems will exhibit an asymmetry of state preparation, if earlier states of the system are causes of later states. Positing an asymmetric causal relation between the states of a system at different times provides an explanation of the asymmetry of state preparation and thus, the asymmetry provides us with an empirically justified reason for modeling physical systems causally, even in the absence of more fine-grained causal relations among the variables characterizing a system's state at different times.

The asymmetry of state preparation often will not be reflected in the dynamical models used to represent a system, but there are cases where the intervention asymmetry plays a direct role in the formalism used to represent a system. Yet again the LHC can provide us with an example of this. We have already discussed the beam control system that measures the state of the proton beam and adjusts the beam in response to these measurements in order to keep the beam parameters stable. One potential problem with feedback systems of this kind is that there is a time lag between the measurements of the beam properties and the response of the control system. Thus, the feedback system would respond to what at the time of intervention is already outdated information and the control would not be optimal. This problem can be addressed by adding a so-called "predictor function" (an example of which is the "Smith predictor"), which uses the measured sate of the beam at some time t to predict its state for a slightly later time $t+\Delta t$ and calibrates the response to the predicted output. Important for our purposes here is that the model of the feedback system is explicitly time-asymmetric: the measurement precedes the intervention into the system by the control mechanism. The model involves a time delay term with a certain sign, about which Steinhagen, for example, says: "Due to the causality, the delay term cannot be inverted." (169)[7]

The temporal asymmetry of causation and intervention plays an important role in the model of the control system, independently of any fine-grained structure of the controlled system—the system intervened upon. Thus, even if Woodward and Hitchcock were right and cetain theories

[7] In a related context, Steinhagen says earlier: "It is obvious that causality forbids the inversion of the exponential delay term." (51)

did not allow us to characterize causal dependencies more fine-grained than between entire time-slices, causal notions can still serve a legitimate role in such theories. But there also ways to introduce causal relata more fine-grained than time-slices.

Woodward worries that if representations in physics would not satisfy condition (IV.3), then his interventionist criterion for causation would not be applicable. The problem is the following. Woodward's definition of total cause specifies that *X* is a total cause of *Y* if and only if the value of *Y* changes under interventions that change the value of *X*. For this definition to be acceptable, the possibility that setting the intervention variable to some value can change the value of *Y* through some route other than through *X* has to be excluded. For example, it has to be excluded that the intervention *I* variable acts as a common cause of *X* and *Y*. Thus, our ability to distinguish between different causal routes is essential to Woodward's interventionist account.

We can avoid this problem, however, if we once again replace Woodward's notion of intervention with Pearl's original notion that directly considers changes in the value of variable intervened upon and does not model these as the result of a causally prior intervention variable. Interventions changing the value of *X* are then associated with changes in the value of some other variable *Y* only if *Y* is causally downstream from *X*—that is, if *Y* is a cause of *X*. If Pearl's stability condition is satisfied, then the intervention condition will be both a necessary and sufficient condition for *X* being a cause of *Y*.

Moreover, there is a mathematical machinery available that allows us to answers questions concerning causal dependencies more fine-grained than that between entire time-slices, even if we are not able adequately to represent a system in terms of a causal structure with discrete nodes linked by discrete causal paths. The machinery I have in mind is that of Green's functions, which can be used to solve inhomogeneous differential equations. The Green's function determines, for example, how placing a small disturbance or a small source of a field at one location changes the field at other locations: the Green's function 'propagates the inhomogeneity,' as it were. Thus, we can model Pearl-style interventions in a field theory as the introduction of an additional field source or as changing the state of a source and then use the Green's functions associated with the source to determine the effect of the intervention.

In this section I have offered two different responses to Woodward's and Hitchcock's worry that an interventionist conception of causation is inapplicable to models that do not allow us to distinguish discrete paths of causal influence. Both of these responses need to be developed further. Do the notions of causation and intervention play a useful role even in contexts in which we do not attribute any causal structure to the system intervened more fine-grained than the state

of a system at a time? As I will argue in chapter 6, the answer to this question is 'yes'. The causal asymmetry characterizing control systems, which I mentioned above, is related to a causal asymmetry at the heart of so-called linear response theory in physics, which models the response of a system to an external force or field—that is, to an intervention into the system. In fact, it is often taken to be one of the major advantages of linear response theory that it allows us to make empirically useful predictions abou the behavior of a system based on very general assumptions—such as a time-asymmetric principle that is usually identified as "principle of causality"—without having to provide a detailed model of the internal structure of the system.

The challenge for my second response is to show how more one can introduce more fine-grained causal structure even for models constructed with the help of theories that, like field theories, do not allow us to distinguish discrete causal paths—and to do so in manner that is compatible with a broadly interventionist conception of causation. In chapter 5 I will show how we can introduce causal structures in the paradigm case of a classical particle-field theory, classical electrodynamics, and argue that these causal structures play both an important inferential and an important explanatory role in the theory. Thus, both cases I will examine meet Woodward's condition of legitimacy, quoted above, according to which "causal notions are legitimate in any context in which we can explain why they are useful, what work they are doing, and how their application is controlled by evidence".

6. Conclusion

In this chapter I have examined various arguments that suggest that causal notions are inapplicable to models constructed with the help of fundamential laws of physics or at the very least that such notions become increasingly strained in the context of physics. The arguments I considered invoke what many take to be the essentially human face of causation. Several such arguments focus on the close relation between the notion of cause and that of intervention or manipulation. But, as I have argued, the notion of intervention can also be usefully employed to models in physics as well. Other features of causal models emphasized by the neo-Russellians—that causal models are partial, incomplete, and involve coarse-graining (in various senses)—are simply features of scientific representations more generally.

In his ground-breaking book on probabilistic theories of causation, Patrick Suppes characterizes contemporary physics this way: "What we are able to get a grip on is a variety of heterogeneous, partial relationships. In the rough and ready sense of ordinary experience, these partial relationships often express causal relations, and it is only natural to talk about causes in

very much the same way that we do in ordinary experience." (Suppes 1970, 6) According to Hitchcock, who endorses what he takes to be Suppes's view, Suppes is arguing that there are preliminary stages in the treatment of phenomena in physics—stages at which these phenomena "still await systematic treatment"—when offering a qualitative causal characterization of the phenomena in question may be legitimate. But as our understanding of the phenomena progresses causal talk eventually ought to be eliminated in favor of a more systematic quantitative treatment. Hitchcock proposes the following principle to capture this view: "(P*) There are advanced stages in the study of certain phenomena when it becomes appropriate to eliminate causal talk in favor of mathematical relationships (or other more precise characterizations)." (Hitchcock 277, 56)

I do not here want to engage in a careful exegesis of Suppes's views, but it seems to me that one could also interpret Suppes's remarks quite differently from the way Hitchcock does, and contrasting my reading with Hitchcock's will allow me to further elucidate the position for which I want to argue here. According to Hitchcock's view and his reading of Suppes, causal characterizations of phenomena are partial, imprecise, merely qualitative, and preliminary. According to my alternative reading, Suppes draws attention primarily to one feature that all representations in physics share with those in the higher sciences and in common sense: their *partiality*. Classical physics may have seduced us with the illusion that physics would ultimately provide us with complete and exact representations of the phenomena, but this was never more than just that—an illusion—and within the partial and heterogeneous representations of the phenomena, which physics actually offers, causal representations continue to play a legitimate and important role. Recall our discussion of the LHC. The models of the proton beam, the accelerator, the particle detectors and of various external influences are partial and are built with an extremely heterogenous range of modeling tools, ranging from elementary particle physics and classical electrodynamics to phenomenological treatments and the mathematics of control systems. The models treat various types of *interventions* into the beams and the construction of some of the models involves what are explicitly identified as time-asymmetric *causal* assumptions. Indeed, as we will see in subsequent chapters, it is precisely the *partiality* of representations, even within the context of classical physics, and the treatment of various *temporal asymmetries* in physics, that provides an argument for the need of causal representations.

But this does *not* mean that causal notions are merely preliminary or provide merely qualitative representations of the phenomena and that we can hope that, in the treatment of a given range of phenomena, these notions will eventually be eliminated. Even though causal models involve coarse-graining, the causal relations posited are not merely qualitative but can be

formulated mathematically precisely. The causal models posited are imprecise only in the same sense in which *all* scientific representations are imprecise. The examples of causal notions in physics which Suppes mentions (and which Hitchcock quotes), support this reading. One such example is dispersion theory and a book edited by Eugene Wigner on this topic: *Dispersion relations and their connection with causality* (Wigner 1964). Dispersion relations occur in disperion theory, which is a special case of linear response theory. While linear response theory in some sense is partial in is description of a system and its response to external forces or fields, the theory and the causal principle at its core are formulated mathematically precisely and not merely in qualitative terms. The connection aluded to in the title of Wigner's book is, as we will see in more detail in chapter 6, a connection between different mathematically precisely formulated principles or relations—a principle of causality from which dispersion relations can be derived. Moreover, it is not obvious, why we should think that linear response theory is waiting to be replaced by some other framework constituting "a more advanced stage in the study" of the phenomena at issue.[8]

The causal principle at the core of linear response theory is a time-asymmetric principle, which invokes explicitly the asymmetry of the causal relation and its connection to the temporal asymmetry. Yet this very asymmetry is frequently taken to provide a powerful argument against the applicability of causal relations to representations in physics. The problem, as it is often presented, is that there is supposed to be a tension between the *causal asymmetry* and the *time-symmetry* or *time-reversal invariance* of the dynamical equations of our fundamental theories. The question whether this contrast can be fashioned into a successful anti-causal argument will be the focus of the next chapter.

[8] To be sure, the quantum-field theoretic S-matrix approach in which dispersion relations occupy center stage has fallen out of favor. But dispersion relations continue to play a role even in more modern treatments of QFT and linear response theory continues to be applied in a wide variety of contexts.