Causality and dispersion: a reply to John Norton

Mathias Frisch
University of Maryland, College Park

Abstract

Classical dispersion relations are derived from a time-asymmetric constraint. I argue that the standard causal interpretation of this constraint plays a scientifically legitimate role in dispersion theory, and hence provides a counterexample to the causal skepticism advanced by John Norton and others. Norton ([???]) argues that the causal interpretation of the time-asymmetric constraint is an empty honorific and that the constraint can be motivated by purely non-causal considerations. In this paper I respond to Norton's criticisms and argue that Norton's skepticism derives its force partly by holding causal principles to a standard too high to be met by other scientifically legitimate constraints.

1. Introduction[1]

Classical dispersion relations are derived from the explicitly time-asymmetric condition that the output field at t_0 is fully determined by the input fields at all times $t \leq t_0$. In (Frisch [???]) I argue that interpreting this condition causally is central to adopting it as physically well-founded. The condition then becomes an expression of the principle that effects do not precede their causes. In his critical reply ([???]) John Norton argues for two claims: first, that the condition should not be interpreted causally and that it 'can and should be founded upon existing electrodynamic theory alone'; and, second, that any attempt to formulate a time-asymmetric causal constraint fails, since the principle either

[1]

does not apply or is too vague to be scientifically legitimate. In this response I want to respond to both claims.

First, however, I want to make a point that might look purely terminological but may in act suggest a disagreement between Norton and me about methodology. The question Norton and I are investigating is whether causal principles are part of classical electrodynamics. But in asking whether the time-asymmetric assumption can be founded upon 'existing electrodynamic theory' Norton already appears to assume what ought to be established by argument, namely that causal principles are not part of 'existing' electrodynamics.

Norton and I agree that the question whether causal notions play a legitimate role in scientific theorizing cannot be addressed by appealing to a priori metaphysics but has to be answered through an examination of actual scientific theorizing. But what counts as legitimate part of theorizing? My assumption is that scientific practice, including how scientists themselves describe the theories and principles at issue ought to be our primary guide in arriving at philosophical accounts of scientific theorizing. Thus, I am inclined to put a lot of weight on the fact that physicists themselves take the time-asymmetric constraint to be a physically well-founded, fundamental causal assumption not in need of further justification. If, as Norton correctly says, we should expect our best science to tell us whether causal notions play a legitimate role in the way we represent the world, then we should also take seriously what scientists take the content of these sciences to be. By contrast, in denying that causal assumption are a part of existing electrodynamics Norton cannot take physicists' appeals to causal notions at face value. Instead he has to explain these appeals away as signs of 'an illusion' that causal notions can provide a principles foundation for the time-asymmetric constraint. Physicists, he is led to suggest, 'succumb to the temptation' of appealing to causal notions as foundation, since it would be 'awkward' for them to admit that the constraint is merely 'opined.'

2. Non-causal foundations?

Norton offers three main reasons in support of his claim that the time-asymmetric condition, which is condition (5) in (Frisch [???]), can be founded upon the purely non-causal content of classical electrodynamics alone:

i) (5) expresses a much more innocuous claim than the causal language in which it is usually couched suggests.

ii) For a simple, special case (5) can actually be deduced from ‘standard electrodynamics’.

iii) General considerations about electrodynamic systems suggest that (5) should be at least in principle derivable from microscopic electrodynamics.

According to Norton, (5) is more innocuous than my causal reading suggests, since it ‘is really only saying that, in the cases we are considering, the dielectric charges respond to incident radiation; they do not anticipate it.’ (Norton [???]) But Norton may here be offering a rephrasing of the condition that is itself causal. The derivation of the dispersion relations begins by positing a general connection between input and output fields that is non-local in time: the output field at x and t depends on the input field at x at all other times (see (2) in Frisch [???]). The question is what reasons we give for assuming that the output field associated with the dielectric varies with, or is responsive to *earlier* input fields but is not responsive to *later* input fields. If we take the notion of *response* to be itself causal, then the condition that a response only occurs after the event to which it responds simply is an instance of the causality condition—however innocuous that assumption may strike us. But if we insist that the notion of response is strictly non-causal, then Norton is proposing an informal rephrasing of the formal constraint (5) without any reasons for why we should accept it. Thus Norton’s formulation either provides a causal reason for the time-asymmetric dependence of the response field on the input field, or it restates the time-asymmetric dependence without offering any reason for its acceptance.

Norton claims that in special cases a derivation of (5) from non causal assumptions can be found in standard textbooks. According to Norton, (Jackson [1999]) deduces ‘the condition [(5)] from standard electrodynamics for a special case (Section 7.10.B) without drawing on causality conditions,’ where ‘standard electrodynamics,’ as Norton emphasizes, is a time reversible theory. But this characterization of Jackson’s derivation is misleading. Jackson does not derive (5) from time-symmetric equations; instead he shows that a particular time-asymmetric model for the dielectric constant ε satisfies the causality condition and then emphasizes that the importance of (5) lies in the fact that it functions as a general constraint with a 'validity that transcends any specific model of ε(ω).'

Jackson also derives the particular model for ε from micro-physical considerations, but this derivation also does not begin with time-reversal invariant laws. Rather, Jackson assumes that the electrons of the medium are bound by a harmonic restoring force acted on by an incoming electric field and subject to a damping force γ. Due to the presence of the phenomenological damping term the equation of motion for an individual electron, with which Jackson begins his derivation, is not time-reversal invariant.

One can show that a damped harmonic oscillator subject to a driving force itself satisfies a causality condition in that its position $x(t)$ time-asymmetrically depends on the driving force at all earlier times (Nussenzveig [1972], sec. 1.2). The dispersion formula derived from this model, according to Nussenzveig, is 'necessarily causal, because it was derived from a causal model.' (46). Thus, the derivation Norton cites as evidence for the claim that the causality condition can be derived from 'standard electrodynamics' does not show that one can recover (5), even for a special case, within a time-reversal invariant microscopic theory of classical electrodynamics. Instead, what it does show is how we can recover the macroscopic time-asymmetric causality condition of dispersion theory from a microscopic time-asymmetric analogue.

Nevertheless Norton is confident that 'there is little doubt that the most general computation, if it could be done, would still return [(5)]' and that 'the causal constraints at issue are merely shorthand for physical constraints already recoverable in classical electrodynamics, though possibly not easily recoverable.' However it is doubtful that the macro-constraint on the interaction between a dispersive medium and fields can in fact be strictly derived from a fundamental classical treatment of the microscopic interactions between charges and fields, since it is doubtful whether there is a fully satisfactory and exact treatment of self-interaction effects—that is, of the interactions of charged particles with their own fields (see Frisch [2005]; [2008]).

Moreover, a derivation of the macro-constraint from a time-reversal invariant micro-theory would show how the putatively causal time-asymmetric constraint reflects an asymmetry characteristic of prevailing initial conditions.[2] In actual systems initial fields are generally not coherently centered on the trajectories of oscillating charges in their future, but the outgoing radiation field of each charge, which is responsible for damped motion, is coherently centered on a point on the trajectory of the charge in the past. The time-reverse process—initial fields 'delicately set up' to converge on the oscillating charges—would presumably result in anti-causal behavior. Importantly, then, in order to show that (5) can indeed be derived from strictly non-causal features of the microscopic physics, one would have to show that this asymmetry characterizing prevailing initial conditions does not itself already reflect causal facts.

In support of his view that no causal facts play a role in explaining the asymmetry in initial conditions Norton points to a tradition that argues that the asymmetry can be explained purely in terms of statistical consideration. A rival view is that elementary classical radiation processes are inherently time-asymmetric. This, for example, appears to have been Albert Einstein's view, who held that in the classical theory 'an oscillating ion produces a diverging spherical wave.' 'The reverse process,' Einstein says, 'does not exist as elementary process. […] The elementary process of the emission of light is, thus, not reversible.' (Einstein [1909], p. 819)[3] On this view the unfamiliar time-reverse of an ordinary radiation process—that is, waves converging into a source—is perfectly possible but unlikely, since it requires carefully set up initial fields that destructively interfere with the diverging radiation fields. By contrast, since oscillating ions produce diverging fields, the familiar radiation damping is what is to be expected.

I argue for a version of Einstein's view in (Frisch [2005]; [2006]), but I do not think that this issue is conclusively settled and certainly cannot be settled here. Let us grant, then, for the moment and for the sake of the argument that a convincing sketch can be given of how macroscopic causal scattering processes arise in systems characterized by certain non-causal asymmetries in prevailing initial conditions. What would follow

[2] Norton maintains that this asymmetry is a matter of choice in that we are free to stipulate the initial conditions, but of course our choice is constrained by our desire to construct models of the kinds of system which we actually, and not as a matter of choice, find in nature.

[3] See (Frisch [2005]) for an attempt to render the view expressed here consistent with the passages from the Einstein-Ritz paper quoted by Norton.

from this for Norton's causal skepticism? It would follow that Norton is correct in claiming that causal notions are not fundamental in the sense that they play no role even in our classical micro-theories of the 'inner constitution of bodies.' But (Norton [2003]) advances several other, stronger theses as well and argues that causal notions belong only to a crude folk science (p. 2) and that such notions are dispensible (p. 8). And neither of these two claims follow from the possibility of an in-principle reduction. Norton himself cites the law of gravitational attraction as an example of an ultimately reducible constraint. But even after the development of the theory of relativity the notion of gravitational force is neither completely dispensable—it is hard to imagine doing the equivalent of classical mechanics without it—nor merely part of a crude folk science. Similarly, even if the asymmetric causal constraint were ultimately in some sense reducible, it remains part of a genuinely scientific theory and within certain contexts is explanatorily indispensable.

3. Other grounds for skepticism

Norton advances several other arguments that are supposed to show independently of the issue of reducibility that there can be no scientifically legitimate principle of causality of the kind I discussed:

(i) I allow for the possibility that the causal principle that effects do not precede their causes is not universal. Norton worries that a sometimes principle is empty and is 'a principle that holds, except when it doesn't.' But this problem—if it is indeed a problem—is faced by every domain-restricted constraint, including, for example, the Maxwell equations themselves. Part of an answer to Norton's worry might be that scientific principles are first posited for a domain as large as possible and that the process of theory acceptance involves both supporting a theory's empirical fruitfulness and establishing restrictions on its domain of validity.

A time-asymmetric causal principle, I suggested, receives broad support from our experimental interactions with experimental systems; and there are formal mathematical frameworks for capturing this asymmetry precisely, as for example (Pearl [2000]) has shown. Norton worries that once we introduce the notion of intervention, the physical systems at issue—target system plus the human intervener—become extremely complicated and it is 'unclear what the intervention experiments reveal.' But intervention experiments are just ordinary experiments, which in the first instance reveal something about the systems intervened on. Even though it seems possible consistently to view experimenters as part of larger asymmetrically causal systems—our experience, after all, is that poking a finger into a light beam affects the beam's propagation at *later* times—there is no general requirement in science that the theory we use to model an experimental system straightforwardly and unproblematically also delivers useful models of the larger system comprising both experimenter and target system.

(ii) Norton claims that considerations involving time-reversed scattering systems allow a reductio ad absurdum of the claim that there is a time-asymmetric causal constraint on dispersive phenomena. He points out that for each scattering process *A* allowed by the Maxwell-Lorentz equations, there is a time-reversed process *B* that is also allowed by the equations. But if one of the two processes satisfies the additional causal constraint, the other process will violate it. So far, however, there no inconsistency. To complete the reductio we have to add as additional premises that (a) the Maxwell-Lorentz equations alone delimit the range of what is physically possible and that (b) systems that violate the causal constraint are physically impossible. Then we can conclude from (b) that the anti-causal system is physically *impossible* and from (a) that the anti-causal system is physically *possible*.

Yet we do not have to accept both (a) and (b). I want to distinguish carefully physical constraints postulated on different levels and note that there are conditions that physicists treat as genuine constraints on what is possible or physically plausible on one level, even though the constraints appear as 'merely' overwhelmingly probable from the perspective of a lower level. Now, the macro-theory of dispersive processes is not time-reversal invariant, since physically plausible models of the dielectric constant, which have to be specified independently in the theory, are not time-symmetric. As I have argued, the causality condition functions as general constraint on plausible models of ε. Thus, on the macro-level (a) is false and the causal constraint together with the Maxwell equations delimits the range of physically plausible scattering processes.

From the perspective of a micro-theory the time-reverse of scattering processes are possible and the time-asymmetric macro-constraints appear merely as overwhelmingly probable. Yet this fact does not render the application of the time-asymmetric *macro*-constraint obscure—the constraint picks out probable familiar scattering phenomena—and it leaves open the question what accounts for the improbability of the unfamiliar, time-reversed micro-processes. On the Einsteinian view that I favor it is precisely the fact that individual charges produce or cause diverging waves which explains why the time-reversed scattering processes are extremely improbable, even though they are possible. On this view, then, (b) is false with respect to constraints on the micro-level, since the causal *micro*-condition does not strictly rule out any temporal evolution allowed by the Maxwell-Lorentz equations but only explains the improbability of certain evolutions. Thus, the reduction goes through neither for a macro-causal constraint nor for a micro-causal condition.

(iii) Norton also argues that there cannot be a universal time-asymmetric causal principle, since, he says, 'there are many cases in which the effect preceding the cause is accepted as a possibility.' But it is important not to conflate different senses of possibility. It may well be that backward causation and closed causal loops are *conceptually* possible or are *possible-according-to-some-theory-T*, but this does not imply that backward causation is *physically* possible or is possible in a universe like ours. The sense in which backward causation is accepted as possibility is perfectly compatible with the existence of an additional time-asymmetric constraint on what kind of causal models adequately represent physical systems in a universe like ours.

4. The principle of energy conservation

Norton contrasts a putative principle of causality with the principle of energy conservation, which he claims is not merely 'decorative' ([2003], p. 3) and is a universal principle 'to which all physical theories must conform' ([2007], p. 231). I want to end this note by pointing out that the roles of the two principles are more closely analogous than Norton allows.[4]

[4] Of course there are also disanalogies between the two principles. My claim here is only that some of Norton's criticisms of the causality condition would seem to apply to energy conservation as well.

Consider how Norton's discussion of the causality condition might be applied to the principle of energy conservation in classical electrodynamics: since the principle follows from the Maxwell-Lorentz equations, it 'can and should be founded upon existing electrodynamic theory alone.' The principle is 'already recoverable in classical electrodynamics' from the fundamental equations, and hence it seems that 'we merely end up assigning an additional adjective ['energy'] to a condition we believe on other grounds.' Moreover, *pace* Norton, the principle is not universal, since, for example, no general principle of energy-momentum conservation can be formulated in General Relativity (see Hoefer [2000]).

Nevertheless, the principle of energy conservation has a legitimate place in physics. Like causal relations, the notion of energy is introduced into a theory's formal framework in a secondary interpretive step, after the basic ontology is fixed. Just as we can interpret the relation between certain variables causally, we identify a certain combination of variables as representing the energy of a system. Arguably there is no 'general property of [a] system that would mark it antecedently' as the system's energy. Rather often it seems one simply looks for some combination of quantities that satisfies an appropriate conservation law.[5] A consequence of the fact that both causal conditions and energy constraints are imposed in a secondary interpretive step is that both types of condition can be applied across wide classes of theories, allow us to unify various models, theories, or frameworks, and can function as heuristic guides in the construction of new theories or models. Moreover, the respective physical interpretations of formal expressions of the two kinds of condition play a crucial role in our applications of the conditions. It seems to be more than an exercise in labeling when we identify a certain formal expression with the energy contained in the electromagnetic field. That a theory satisfies a principle of energy conservation is treated as a theoretical desideratum and the fact that the theory allows us to formulate such a principle increases our confidence in the theory. Similarly, that a model of ε satisfies condition (5) can be understood as a desideratum on all such models precisely because (5) is interpreted as a general causal constraint. Thus, as I argue in (Frisch [???]), both energy conservation and the causality

[5] See (Parrott [1987]) for a survey of several different proposals for what expression should be identified with the energy of the electromagnetic field.

condition are general principles (as opposed to ‘mechanism theories’) in a broadly Lorentzian sense.

Nobody would doubt that principles of energy-momentum conservation play a legitimate role in science. The fact that some of Norton’s argument aimed at relegating causal conditions to the status of a mere folk science seem to apply to energy conservation with equal force suggests that he holds causal principles to a standard too high for actual science.

Acknowledgements

This project was partially funded by the National Science Foundation, award no. SES-0646677.

References

Einstein, A. [1909]: 'Über die Entwicklung unserer Anschauung über das Wesen und die Konstitution der Strahlung', *Physikalische Zeitschrift* (**10**), pp. 817-25.

Frisch, M. [???] '“The most Sacred Tenet”? Causal Reasoning in Physics', *British Journal for the Philosophy of Science*.

---. [2005]: *Inconsistency, Asymmetry, and Non-Locality: A Philosophical Investigation of Classical Electrodynamics*. Oxford: Oxford University Press.

---. [2006]: 'A tale of two arrows', *Studies In History and Philosophy of Modern Physics* **37** (3), pp. 542-58.

---. [2008]: 'Conceptual Problems in Classical Electrodynamics', *Philosophy of Science* **75** (1), pp. 93-105.

Hoefer, C. [2000]: 'Energy Conservation in GTR', *Studies In History and Philosophy of Modern Physics* **31** (2), pp. 187-99.

Jackson, J. D. [1999]: *Classical Electrodynamics*. 3rd ed. New York: Wiley.

Norton, J. D. [???] 'Is There an Independent Principle of Causality in Physics?' *British Journal for the Philosophy of Science*.

---. [2003]: 'Causation as Folk Science', *Philosophers' Imprint* 3 (4), http://www.philosophersimprint.org/003004/ 3, pp. 1-22.

---. [2007]: 'Do the Causal Principles of Modern Physics Contradict Causal Anti-Fundamentalism?' In P. K. Machamer and G. Wolters (eds.), 2007, *Thinking about Causes: From Greek Philosophy to Modern Physics*, Pittsburgh: University of Pittsburgh Press.

Nussenzveig, H. M. [1972]: *Causality and Dispersion Relations*. New York: Academic Press.

Parrott, S. [1987]: *Relativistic Electrodynamics and Differential Geometry*. New York: Springer-Verlag.

Pearl, J. [2000]: *Causality: Models, Reasoning, and Inference*. Cambridge, U.K: Cambridge University Press.