Philosophy of Physics

How Oriented Causation Is Rooted into Thermodynamics

CARLO ROVELLI

RESEARCH

LSE Press

ABSTRACT

The notions of *cause* and *effect* are widely employed in science and commonly understood as temporally ordered: causes precede effects. I discuss why and how these notions are rooted into thermodynamics. The entropy gradient (i) explains in which sense interventions affect the future rather than the past, and (ii) underpins the time orientation of the subject of knowledge as a physical system. Via these two distinct paths, it is the entropy gradient, and only the entropy gradient, the source of the time orientation of causation, namely the fact the cause comes *before* its effects.

CORRESPONDING AUTHOR:

Carlo Rovelli

Aix-Marseille University, Université de Toulon, CPT-CNRS, F-13288 Marseille, France; Department of Philosophy and the Rotman Institute of Philosophy, 1151 Richmond St. N London N6A5B7, Canada; Perimeter Institute, 31 Caroline Street N, Waterloo ON, N2L2Y5, Canada

rovelli.carlo@gmail.com

TO CITE THIS ARTICLE:

Rovelli, Carlo. 2023. "How Oriented Causation Is Rooted into Thermodynamics." *Philosophy of Physics* 1(1): 11, 1–14. DOI: https://doi.org/10.31389/pop.46

DOI: 10.31389/pop.46

cause of the destruction of Pompei, my push is the cause of the door opening, shots by Lee Oswald caused the death of John Kennedy, smoking is a cause of cancer, a stone falling into a pond causes waves, and so on. I start with examples rather than a definition because clarifying which definition is implicitly utilized when we talk about these causes is one of the aims of what follows.

Among scientists working in fundamental theoretical physics, it is commonly assumed that causation (in the sense of these examples) does not play any role in the elementary physical description of the world. In fact, no fundamental elementary law describing the physical world that we have found is expressed in terms of causes and effects. Rather, laws are expressed as regularities, in particular describing correlations, among the natural phenomena.[1]

Furthermore, in contemporary fundamental physics the elementary laws that describe nature are expressed by correlations that do not distinguish past from future: they do not have any orientation in time.[2] Hence they alone cannot imply any time-oriented causation. This fact was emphasized by Bertrand Russell (1913), who opens his influential article *On the Notion of Cause*, claiming that "'cause' is so inextricably bound up with misleading associations as to make its complete extrusion from the philosophical vocabulary desirable."[3]

However, the usage of the notion of cause is far from fading, neither from the philosophical nor from the scientific language, and is definitely distinguished from the notion of correlation (Cartwright 2007). In science, causal explanations are not only ubiquitous, they also play an essential role, distinct from correlations: smoking and entering a hospital are both correlated with death by lung cancer, but smoking causes cancer, entering the hospital doesn't.

Cartwright (2007) argued that the fact that causation is ubiquitous in science implies that it must be a universal feature of nature, and that its absence in fundamental physics

---

1 By "causation" I mean here what is called "strong causation" in (Adlam 2022). This is the oriented notion of causation of our intuition, where *A* causes *B* and not vice versa. This should not be confused with relativistic causality, namely the fact that that there cannot be space-like influences. The first is asymmetric, the second symmetric. In quantum field theory, the second is expressed by the fact that operators supported in space-like separated regions commute. To be "causally connected" in this symmetrical sense is a much weaker notion. And so is the more general symmetric notion of causal connection considered in (Adlam 2022), analyzing nonstandard quantum causal models. To be causally connected (in this weaker sense, or in the sense of relativistic causality) is a condition for causation to be possible. Strong causation is much more than this, as clarified below, and is directional.

2 A mistake in the literature is to state that parity-charge violation and CPT invariance determine an absolute preferred time direction. This is wrong. What to call left/right direction, or positive/negative charge is conventional; hence nothing measurable picks a preferred absolute time direction among two histories related by CPT. This is the same mistake as the claim that the Maxwell equations pick a direction of time because a time-reversed solution is a solution only by flipping the sign of the electric field. It stems from misunderstanding what is the puzzling aspect of the lack of *relevant* distinction between past and future.

3 The idea that causation is nothing other than correlation and that the distinction between "cause" and "effect" is nothing other than the distinction between what comes first and what comes next in time can be traced to Hume, for whom causation is correlations between contiguous events. But Hume is actually subtler in the Treatise: He identifies causation not with the correlation itself, but with the idea in the mind that is determined by noticing these correlations. "An object precedent and contiguous to another, and so united with it, that the idea of the one determines the mind to form the idea of the other, and the impression of the one to form a more lively idea of the other" (Hume 1736). Even more so in the Enquiry: "custom … renders our experience useful to us, and makes us expect, for the future, a similar train of events with those which have appeared in the past" (Hume 1777). I will come back to this point in Section IV.

testifies for the disunity of science. I do not think that either follows. Rather, causation is an important concept that we use widely and appropriately, even if it does not make sense at the elementary physical level. Many concepts are like that.[4]

This consideration raises a question: what do we precisely mean by oriented causation, given that there is no oriented causation at the elementary physical level? This is what I discuss here. Clarifying this point contributes to the account of the unity of science, not its disunity.

III. CAUSAL MODELS

Light has been shed on what we mean by causation in this sense by the introduction of causal modeling techniques (Hitchcock 2023). These provide concrete working algorithms for discovering relations that are specifically causal (for instance, distinguishing smoking from hospitalization as causes for cancer), hence making the notion of causation precisely defined.

The way this works can be briefly synthetized as follows. Assume we have a number of variables A, B, C, D, … that are partially ordered in time, and we know—experimentally or theoretically—that the values that these variables can take are correlated. Assume we know these correlations. Then we say that the variable B is causally related to a variable C that follows it in time if the following happens. If we disregard the correlations between B and *earlier* variables, that is, we allow these correlations to violated, and *intervene* setting the variable of B at different values, then the (known) correlations in the future of B imply that differences in the value of B affect the value of C.

In the smoking example, this signifies that "smoking is a cause for cancer" and "going to the hospital is not" *means* that if we intervene by preventing people from smoking, then we expect the incidence of lung cancer to decrease, but not so if we prevent people from going to the hospital. This is true, irrespectively from the fact that both are positively correlated with cancer.

The clarification of the notion of causation brought by the causal modeling techniques is based on two ingredients. The first is the notion of intervention. The second is the idea that intervention affects the future, not the past. I consider the two points in detail, below.

A. INTERVENTION

In the causal models, intervention is an interaction between the set of variables considered, A, B, C, D, … and *something else*, which upsets the relations between these variables coded in the correlations they have when there is no intervention. What does this something else represent?

In the intuition at the basis of some of these models and in many applications, intervention describes a manipulation by a human agent: I do something and my action causes an effect. But the notion of intervention does not require this anthropocentric interpretation. We say that a meteorite falling on the moon causes the formation of a crater, or that intense volcanic activity during a certain geological era is the cause of the presence of a certain chemical in a geological stratum at later times. In these cases, as in many others, the agent intervening is not human, and not biological. It can be a meteorite or volcanic activity.

---

4 Energy conservation, for instance, does not always make sense when relativistic gravity is involved.

What is sufficient to define the notion of an intervention is the fact that a certain system (the surface of the moon, the surface of the earth, the door opening, Pompei, Kennedy, the body, the pond …) whose dynamics and correlations we understand, interacts with another system, whose dynamics we disregard, and which we treat as an external agent: the meteorite, the volcanic activity, the shots by Lee Oswald, something preventing smoking, the fall of the stone, etc. The causal models treat the system as a set of correlated variables, but they treat the agent-system as acting arbitrarily, or "freely"; that is, they disregard its dynamics. This will prove crucial in understanding causation.

B. TIME ORIENTATION

In causal modeling, the assumption that breaks time reversal invariance is the assumption that intervention of the external agent affects the future, not the past. It is important to notice that this is *assumed* in the definition of causation; intervention is assumed to break *past* correlations, while *future* ones are preserved and determine the effect.[5]

The question is: why do we assume that *past* correlations are broken, if the elementary laws of nature are time reversal invariant and do not make any distinction between the past and the future direction of time?[6] The reason is obvious—the laws of nature are time reversal invariant, but in the world around us, agency does nevertheless affect the future, not the past. What does this imply?

Instead of considering how the future would change if the past was the same and the intervention had not occurred, we could equally consider how the past would change if the future was the same and the intervention had not occurred. That is, we could consider the question of what would happen to the past if we cut ties to future variables and asked how present interventions would affect the past.[7] Yet, such reverse logic does not work with causation. Why? What breaks the time reversal symmetry?

To see what breaks the time reversal symmetry, and understand where the time orientation is coming from, consider an example: the history of a pond hit by a stone at a certain location O at time $t = 0$. Round concentric waves form around O after the impact. They move outwards and agitate the pond for a while, until their energy dissipates and the pond goes back to a quiet equilibrium state. Let's treat the fall of the stone as an external intervention, that might or might not have occurred. Consider the case in which the stone did not fall, and ask what would have happened to the pond, according to the physics we know. To determine a history, physical laws need the state of the system to be specified at some time. On the basis of this, they determine the history of the system *in both time*

5 The violation of dynamics in the immediate past of the intervention plays the role of the "tiny miracle" in Lewis's account of counterfactual dependence. Notice that it does not require violations of physical laws. It is the comparison of two possibilities—two "worlds" in Lewis's terminology—which can both satisfy the physical laws of the actual world, one in which the agent intervenes and one in which it does not. In the world where the agent intervenes, what are broken are the correlations that describe the system *when not interacting*, and not the physical laws of the world, which of course include the agent.

6 For an arbitrary invertible evolution, dependence may not be symmetric. For instance, for $f : (A, B) \to (C = A, D = A + B)$, the variables A and D are dependent, but for the inverse evolution $f^{-1} : (C, D) \to (A = C, B = D - C)$, they are not. This does not happen in classical and quantum mechanics: f is not time-reversal invariant. Thanks to Robert Spekken for making me notice this point.

7 We do use such past counterfactuals—the light is on, and you tell me that you have just turned it on; if you hadn't, the light being on would have implied that we left it on earlier, when we last left home. In these cases, we consider different pasts with the same future, determined by the intervention.

directions. So we have a choice—we can ask what would have happened to the pond if the stone had not fallen:

(a) if the history of the pond had been the same as in the history with the falling stone for all times $t < 0$, or

(b) if the history of the pond had been the same as in the history with the falling stone for all times $t > 0$.

Physics answers in both cases. In both, the answer is a physical history which is consistent with elementary mechanics. But the two cases are nevertheless remarkably different in our experience:

(a) In the first, nothing remarkable happens for $t > 0$. The surface of the pond remains quiet, instead of being excited in concentric waves.

(b) In the second, we have to find a past history h of the pond (never hit by anything) such that from time $t = 0$ onward there are spherical outgoing waves from a certain location. This history certainly exists (the water of the pond might have moved in all sort of ways in the past), but it looks strangely implausible to us. Why so? Because we know from experience that the water of the pond tends to equilibrate and go back to an equilibrium state. How come the water of the pond is still agitated in this strange and unnatural manner, if nothing had happened?

Thus, among the possible alternatives that we can consider could have happened if the intervention had not occurred, only the ones *with the same past* are compatible with the time-oriented world in which we happen to live. (That world is "closer to actuality" (Lewis 1973).)

The key point is that this argument is not about the mechanical laws of physics: These are compatible with the history h. It is something else that makes h strange. What is it? It is thermodynamics: we know that systems equilibrate in time. Thermodynamics has more ingredients than elementary physical laws: a macroscopic description of the world and the observation that entropy was low in the past.[8]

A thermodynamic account of the intervention, in fact, clarifies the source of the asymmetry in time. The pond is in a near equilibrium state a certain temperature T. The stone that hits the pond has a kinetic energy E much larger than the average kinetic energy per degree of freedom of the water molecules, namely $E \gg kT$, where k is the Boltzmann constant. Hence, the stone is out of equilibrium with the pond. Hence thermodynamics predicts (probabilistically!) that energy is transferred from the stone to the pond. More precisely, at time $t = 0$, energy is transferred to water molecules in the vicinity of O. The transferred energy is free energy for the pond—it is the energy of the outgoing concentric waves– and thermodynamics tells us that (probabilistically!) this free energy is going to be dissipated into the water, raising its temperature and moving toward energy equipartition.

In Lewis's terminology, the world without intervention w_o can be compared with a world with intervention w_p having the same past as w_o, or a world with intervention w_f having the same future as w_o. Both w_p and w_f satisfy the dynamical laws, but w_p satisfies also the second law of thermodynamics (and its generalizations) while w_f does not. Hence, w_p is a

8 Here, by "thermodynamics," I also mean the tendency of systems to equilibrate toward the future, following past lower entropy. Some authors (correctly) distinguish thermodynamics from equilibration (Brown and Uffink 2001; Myrvold 2020).

DOI: 10.31389/pop.46

world far closer to ours (again in Lewis language) than w_f. This is why we talk about w_p as a reasonable possibility and not so about w_f. And this is therefore why we consider the effect to be in the future of the cause.

This account makes clear why the effect of the intervention is in the future of the intervention and not in its past: Because it is in the past that the overall system was away from equilibrium, the stone had more energy than what equipartition of energy indicates. Because of this past low entropy, the thermodynamical evolution of the system is time-oriented, as for all thermodynamical systems with past low entropy.

Thus, the effect (the waves) *follows* the cause (the impact of the stone) instead of preceding it because of the time orientation given by the existence of a thermodynamic disequilibrium *in the past*.

The waves, namely the effect, follow the fall of the stone, namely the cause, because it is a step in the thermodynamic equilibration of an initial unbalance.

A different answer considered in the literature derives the arrow of causation from the epistemic arrow, and this from the asymmetry of records (Albert 2000; 2015). Causes precede effects because they are fixed, while the effect is not; they are fixed because they are in the past and we consider the past fixed because we have traces and memories of it. I am not convinced this is sufficient to account for the time-oriented notion of causation. There are cases in which we know about the future and we do not know about the past, and still we do not say that a cause in the future has an effect in the past. For instance, if tomorrow we discover a fossil, we learn about the past, but in this case, we do not say the fossil is the cause of the extinct animal: we only say that it is the cause of our knowledge of the extinct animal.

Still, let me accept here that the asymmetry of causation can be reduced to the epistemic time arrow. Would this change the general conclusion? I do not think so, for the following reason: As discussed in detail in (Rovelli 2022a), to which I refer for an extensive analysis of the question, traces, and memories of the past are only possible because we live in an entropy gradient. At equilibrium, there is no sense in which traces and memories distinguish past from future. Hence the root of the asymmetry is once again the contingent entropy gradient.

C. THE THERMODYNAMIC ARROW IS ESSENTIAL

The structure of the relation between what we call a cause and what we call its effect illustrated above with the example of the stone in the pond is generic: its time orientation is given by the thermodynamic arrow of time, namely the actual fact that entropy was lower in the past. Causation is therefore a macroscopic thermodynamic phenomenon where the total entropy is raised by an intervention, and the effect is the trace left on the system by the intervention.

This thermodynamic structure is the same as that characterizing *traces* in general and *memory* in particular (Rovelli 2022a). A trace or a memory are indeed effects of the event they record, which is their cause. This is also the general structure of the phenomenon we call "agency" (Rovelli 2021). More on this later.

Can this be proven in general? Yes, and the proof is surprisingly simple. Precisely because causation does capture a time-oriented relation, it can only depend on the single source of time orientation that is compatible with physics: the thermodynamic arrow of time.

Namely, the fact that we live in a universe that we describe macroscopically and which has a consistent temporal orientation of its entropy gradients.[9]

To further clarify this fact, consider the two extreme situations where there is no entropy gradient. The first is when the overall system is at or near thermal equilibrium. In this case, the energy of the stone falling into the pond must be of the order kT. If so, its effect on the water molecules is indistinguishable from the generic thermal agitation: the fall has no detectable effect. There is no time orientation and no sense of causation.

The other case is the case of purely mechanical interactions, without any coarse-grained description. In this context, there is no notion of entropy, no notion of more or less probable macroscopic states, and therefore no sense in which considering histories with the same future can be less plausible than considering histories with the same past. There is no intrinsic distinction between past and future and therefore, again, no possible intrinsic time-oriented causation. If two stones with velocities v_1 and v_2 collide and after the collision have velocities w_1 and w_2, then there is nothing in the phenomenon itself that fixes a preferred time orientation. We can equally say that without the collision, the velocities would have been always v_1 and v_2 or always w_1 and w_2. If we say that the *later* velocities are caused by the collision, we are truly in the situation described by Hume: we notice a correlation and the term we call "cause" is only characterized by the fact that it happens earlier. Causation in this sense is reduced to (symmetric) correlation.

This seems to exhaust the analysis of the source of the time orientation of causation. Is that all? No. I believe we are are still missing *the* crucial ingredient of this story.

IV. THE AGENT'S TIME ARROWS

If the above is physically correct, why is it of any relevance? As defined above, causation is an intricate and baroque thermodynamic phenomenon describing certain peculiar statistical patterns in the interaction between systems. Why is then causation so important in our making sense of the world?

As Huw Price puts it (tracing the idea to Ramsey (1978)): The interesting question isn't what causation is, but how we come to think and talk in causal terms (Price 2023). In other words, we do not understand causation by asking which complicated patterns in the tapestry of nature we happen to call causation; we understand causation by understanding why those patterns are relevant at all, which is to say, by understanding what is causal thinking.

In the previous pages, there were numerous hints about the answer: (a) the intuition about the notion of intervention comes from *our own* capacity of intervening and manipulating systems; (b) we are interested in the fact that it is smoking that causes cancer because *we can actually intervene on this*; (c) as Hume points out, causation is not really something happening in the phenomena—it is *the idea in the mind that we gather from that which is useful to us*. All this points to the fact that we use causation as a predictive tool for possible futures, where *we ourselves are the agents* that can intervene.

We are subjects of knowledge and actors in the world that have a direct involvement in the causation game and a direct interest in causal relations. After all, our brain is essentially a

9 Attempts to trace the arrow of time to something else (such as Gold (2005)) are either unconvincing or can themselves be traced to the statistical irreversibility connected to low entropy in the past. On perspectival grounding for irreversibility (Price 2023), see the discussion below, in Section IV.

machine that analyses the different possible futures that would follow *if this or that course of action is taken*. The main business our brain is involved in is not simply to predict the future given the past (Buonomano 2017), but to predict what would happen under different choices of behavior, namely to predict what would the effects of different interventions be.

A (neo-)pragmatist perspective is therefore a natural context to understand why causation is key to our understanding of the world. That is, pragmatism is particularly convenient for explaining the relevance of causation, and what causation is, in the perspective of the subject, as emphasized by Huw Price (2023; 2007). Citing Ramsey again, "from the situation when we are deliberating seems to me to arise the general difference of cause and effect" (Ramsey 1978).

At the light of this insightful consideration, causality is better understood, I believe, not as a simple physical issue, and even less as a metaphysical one, but rather in terms of the perspective and interest of deliberating human agents. This is in tune with Cartwright's observation that a crucial role of causal notions is distinguishing effective from ineffective strategies (Cartwright 1979).

However, if we do not want to give up naturalism, we cannot treat the subject as unphysical. The subject may be peculiar in various manners but we expect it to be also a physical system like any other physical system. It is itself a natural being subjected to the laws of physics. Hence we can investigate the agent that utilizes causation in the way we have discussed, which determines its relevance. Its peculiar behavior must be grounded into physics. In particular, its own time orientation cannot but itself be grounded in physics. Let's analyze this fact in detail. The following are facts:

(a) The subject knows the past, more than the future.

(b) The subject can choose which actions to do.

(c) Its choice affects the future, not the past.

The first is the epistemic arrow of time (we know the past better than the future). The second is the vivid intuition of our freedom in choosing, which sometimes goes under the name of free will. The third is the agential arrow of time (we can affect the future, not the past). Crucially, the time orientation of these aspects of the subject's phenomenology are themselves effects of the thermodynamic arrow of time.

The epistemic arrow of time is a consequence of the fact that an entropy gradient plus some additional simple conditions largely realized in our universe (long thermalization times and systems' separation) are sufficient for the formation of abundant traces of the past (that have no time reversed equivalent). The past is fixed because in the present there are traces about it, and these are there because of entropy was low in the past. This was anticipated in Reichenbach (1956) and is discussed in detail in Rovelli (2022a).

The possibility for the agent of determining different macroscopic futures is also permitted by an entropy gradient. There is a simple way to prove this (Rovelli 2022b). A choice is a (physical) process which macroscopically can evolve into two different futures (with the same past). This is not in contradiction with determinism (Spinoza 1677) because it is a macroscopic description: different micro-histories that in the past were part of the same macro-state can evolve into different macro-states (Loewer 2007; 2012; 2020; Rovelli 2021).

If we look at the macrophysics only, there is a gain of information at the choice—the information about which choice was actualized. This new macroscopic information is generally paid for in entropy increase, that is, dissipation (which loses macroscopic

information). Agency can therefore be interpreted as a macroscopic thermodynamic phenomenon that gets its time orientation from the entropy gradient and dissipation. It is in this macroscopic context that the notion of causation acquires salience, the context where thermodynamics is relevant and time orientation makes sense.

Now, intervention was defined by assuming the intervening agent to be "free": This is possible precisely because it amounts to disregarding some degrees of freedom and their dynamics. Agency is grounded in the ignoring of degrees of freedom and their dynamics (Spinoza 1677).

Biological agents, in particular, determine their own behavior in part by calculating future outcomes of their possible alternative. The embedded (physical!) information they utilize is itself part of what determines the future and as such it interferes (in the sense of Ismael (2022; 2023)) with what is going to happen. Agents (in this sense) necessarily work with incomplete information. For the agent's (embedded, physical) representation of the world, the future is therefore necessarily "open"; it depends on its choices (Ismael 2022; 2023).

The alternative perspective, which takes into account how come we do so, is more clarifying: The entropy gradient of the world creates opportunities for the use of information to guide behavior. Hence we have evolved to pick out from the environment macroscopic variables to interact with, suitable for us to act upon the world in ways that we can (partially) causally control to our own advantage (Ismael 2022; 2023).

The question of *what is it in the natural world that enables us (humans or generic organisms) to use this opportunity* is a well posed issue, and a scientific one. The answer is the entropy gradient. From the physical perspective, all the features of the subject that motivates its interest in causation are therefore *themselves* rooted in the thermodynamic arrow of time.

In a slogan, In thermal equilibrium there is no agent interested in causation.

In closing, let's get back to the situation where we interpret correlation as causation in settings where there is no dissipation, namely no thermodynamic time orientation in the phenomena themselves. For instance, in the case of two particles colliding, for which the velocities were v_1 and v_2 before the collision and are w_1 and w_2 after the collision. In this case, we *still* say that the collision *caused* the velocities to become w_1 and w_2. Not (as argued above) because of an intrinsic time orientation in the particle's phenomenology, but just because of the time orientation of us subjects, itself rooted in the entropy gradient that nourishes us. (On this, see also Section 5 of Price (2007).) Causation is salient for a deliberating agent whose time orientation comes from the entropy gradient.

V. CONCLUSION

According to current physics, the basic laws of our universe are time reversal invariant—they do not determine a preferred direction of time. They express correlations between events, which are symmetric in time. There is no notion of causation at this level. This came as a historical surprise.

On the other hand, the history in which the universe happens to be is not symmetric under time reversal. We access it only via a relative small number of degrees of freedom, and the entropy that such coarse graining defines happened to be far lower in a direction of time, the one we denote as past. There is a commonly oriented entropy gradient in all phenomena we witness, as well as in our own behavior.